Beijing 2002
August 20–28

Proceedings of the

International Congress of Mathematicians

Vol. II: Invited Lectures

Higher Education Press

World Scientific
www.worldscientific.com

International Congress of Mathematicians (2002, Beijing)
Proceedings of the International Congress of Mathematicians
August 20–28, 2002, Beijing

Editor: LI Tatsien (LI Daqian) dqli@fudan.edu.cn

Editorial Assistants: Cai Zhijie, Lu Fang, Xue Mi, Zhou Chunlian

This volume is the first part of the collection of manuscripts of the lectures given by the invited speakers of the ICM2002. The second part of this collection is published in Volume III.

The manuscripts of the invited lectures are ordered by sections and, in each section, alphabetically by author's names. In case of several authors for one manuscript, the name of invited speaker is written in boldface type.

The electronic version of this volume will be published on the international Math ArXiv with the address

http://front.math.ucdavis.edu/

55 Shatan Houjie, Beijing 100009, China
http://www.hep.com.cn http://www.hep.edu.cn

Copy Editors: Li Rui, Li Yanfu, Wang Yu

ISBN 7-04-008690-5 Set of 3 Volumes

Contents

Section 1. Logic

E. Bouscaren: *Groups Interpretable in Theories of Fields* 3

J. Denef, F. Loeser: *Motivic Integration and the Grothendieck Group of Pseudo-Finite Fields* 13

D. Lascar: *Automorphism Groups of Saturated Structures; A Review* 25

Section 2. Algebra

S. Bigelow: *Representations of Braid Groups* 37

A. Bondal, D. Orlov: *Derived Categories of Coherent Sheaves* 47

M. Levine: *Algebraic Cobordism* 57

Cheryl E. Praeger: *Permutation Groups and Normal Subgroups* 67

Markus Rost: *Norm Varieties and Algebraic Cobordism* 77

Z. Sela: *Diophantine Geometry over Groups and the Elementary Theory of Free and Hyperbolic Groups* 87

J. T. Stafford: *Noncommutative Projective Geometry* 93

Dimitri Tamarkin: *Deformations of Chiral Algebras* 105

Section 3. Number Theory

J. W. Cogdell, I. I. Piatetski-Shapiro: *Converse Theorems, Functoriality, and Applications to Number Theory* 119

H. Cohen: *Constructing and Counting Number Fields* 129

Jean-Marc Fontaine: *Analyse p-adique et Représentations Galoisiennes* 139

A. Huber, G. Kings: *Equivariant Bloch-Kato Conjecture and Non-abelian Iwasawa Main Conjecture* 149

Kazuya Kato: *Tamagawa Number Conjecture for zeta Values* 163

Stephen S. Kudla: *Derivatives of Eisenstein Series and Arithmetic Geometry* 173

Barry Mazur, Karl Rubin: *Elliptic Curves and Class Field Theory* 185

Emmanuel Ullmo: *Théorie Ergodique et Géométrie Arithmétique* 197

Trevor D. Wooley: *Diophantine Methods for Exponential Sums, and Exponential Sums for Diophantine Problems* 207

Section 4. Differential Geometry

B. Andrews: *Positively Curved Surfaces in the Three-sphere* 221
Robert Bartnik: *Mass and 3-metrics of Non-negative Scalar Curvature* 231
P. Biran: *Geometry of Symplectic Intersections* 241
Hubert L. Bray: *Black Holes and the Penrose Inequality in General Relativity* 257
Xiuxiong Chen: *Recent Progress in Kähler Geometry* 273
Weiyue Ding: *On the Schrödinger Flows* 283
P. Li: *Differential Geometry via Harmonic Functions* 293
Yiming Long: *Index Iteration Theory for Symplectic Paths with Applications to Nonlinear Hamiltonian Systems* 303
Anton Petrunin: *Some Applications of Collapsing with Bounded Curvature* 315
Xiaochun Rong: *Collapsed Riemannian Manifolds with Bounded Sectional Curvature* 323
Richard Evan Schwartz: *Complex Hyperbolic Triangle Groups* 339
Paul Seidel: *Fukaya Categories and Deformations* 351
Weiping Zhang: *Heat Kernels and the Index Theorems on Even and Odd Dimensional Manifolds* 361

Section 5. Topology

Mladen Bestvina: *The Topology of Out(F_n)* 373
Yu. V. Chekanov: *Invariants of Legendrian Knots* 385
M. Furuta: *Finite Dimensional Approximations in Geometry* 395
Emmanuel Giroux: *Géométrie de Contact: de la Dimension Trois vers les Dimensions Supérieures* 405
Lars Hesselholt: *Algebraic K-theory and Trace Invariants* 415
Eleny-Nicoleta Ionel: *Symplectic Sums and Gromov-Witten Invariants* 427
Peter Teichner: *Knots, von Neumann Signatures, and Grope Cobordism* 437
Ulrike Tillmann: *Strings and the Stable Cohomology of Mapping Class Groups* 447
Shicheng Wang: *Non-zero Degree Maps between 3-Manifolds* 457

Section 6. Algebraic and Complex Geometry

Hélène Esnault: *Characteristic Classes of Flat Bundles and Determinant of the Gauss-Manin Connection* 471
L. Göttsche: *Hilbert Schemes of Points on Surfaces* 483
Shigeru Mukai: *Vector Bundles on a K3 Surface* 495

R. Pandharipande: *Three Questions in Gromov-Witten Theory* 503
Miles Reid: *Update on 3-folds* 513
Vadim Schechtman: *Sur les Algèbres Vertex Attachées aux Variétés Algébriques* 525
B. Totaro: *Topology of Singular Algebraic Varieties* 533

Section 7. Lie Group and Representation Theory

Patrick Delorme: *Harmonic Analysis on Real Reductive Symmetric Spaces* 545
Pavel Etingof: *On the Dynamical Yang-Baxter Equation* 555
D. Gaitsgory: *Geometric Langlands Correspondence for GL_n* 571
Michael Harris: *On the Local Langlands Correspondence* 583
Alexander Klyachko: *Vector Bundles, Linear Representations, and Spectral Problems* 599
Toshiyuki Kobayashi: *Branching Problems of Unitary Representations* 615
Vikram Bhagvandas Mehta: *Representations of Algebraic Groups and Principal Bundles on Algebraic Varieties* 629
E. Meinrenken: *Clifford Algebras and the Duflo Isomorphism* 637
Maxim Nazarov: *Representations of Yangians Associated with Skew Young Diagrams* 643
Freydoon Shahidi: *Automorphic L-Functions and Functoriality* 655
Marie-France Vignéras: *Modular Representations of p-adic Groups and of Affine Hecke Algebras* 667

Section 8. Real and Complex Analysis

A. Eremenko: *Value Distribution and Potential Theory* 681
Juha Heinonen: *The Branch Set of a Quasiregular Mapping* 691
Carlos E. Kenig: *Harmonic Measure and "Locally Flat" Domains* 701
Nicolas Lerner: *Solving Pseudo-Differential Equations* 711
C. Thiele: *Singular Integrals Meet Modulation Invariance* 721
S. Zelditch: *Asymptotics of Polynomials and Eigenfunctions* 733
Xiangyu Zhou: *Some Results Related to Group Actions in Several Complex Variables* 743

Section 9. Operator Algebras and Functional Analysis

Semyon Alesker: *Algebraic Structures on Valuations, Their Properties and Applications* 757
P. Biane: *Free Probability and Combinatorics* 765
D. Bisch: *Subfactors and Planar Algebras* 775

Liming Ge: *Free Probability, Free Entropy and Applications to von Neumann Algebras* 787
V. Lafforgue: *Banach KK-theory and the Baum-Connes Conjecture* 795
R. Latała: *On Some Inequalities for Gaussian Measures* 813

Author Index 823

Section 1. Logic

E. Bouscaren: *Groups Interpretable in Theories of Fields* 3

J. Denef, F. Loeser: *Motivic Integration and the Grothendieck Group of Pseudo-Finite Fields* 13

D. Lascar: *Automorphism Groups of Saturated Structures; A Review* 25

ICM 2002 · Vol. II · 3–12

Groups Interpretable in Theories of Fields

E. Bouscaren*

Abstract

We survey some results on the structure of the groups which are definable in theories of fields involved in the applications of model theory to Diophantine geometry. We focus more particularly on separably closed fields of finite degree of imperfection.

2000 Mathematics Subject Classification: 03C60, 03C45, 12L12.
Keywords and Phrases: Groups, Fields, Definability, Algebraic groups.

1. Introduction

In the last ten years, the model theory of fields has seen striking new developments, with applications in particular to differential algebra and Diophantine geometry. One of the main ingredients in these applications is the analysis of the structure of groups definable in fields with added "definable structure".

Model theory studies structures with a family of distinguished subsets of their Cartesian products, the family of *definable* subsets, which is requested to be closed under finite Boolean operations and projections. In the case of algebraically closed fields, the definable sets are exactly the constructible sets in the Zariski topology (finite Boolean combinations of Zariski closed sets). If one considers fields which are not algebraically closed (for example, fields of positive characteristic which are separably closed and not perfect) or algebraically closed fields with new operators (differentially closed fields, fields with a generic automorphism), then the family of definable sets is much richer than the family of Zariski constructible sets. In each of the above cases, one can generalize the classical geometric notions, by using the tools developed by model theory (abstract notion of independence, of dimensions...). For example:

1. One can define "good" topologies which strictly contain the Zariski topology.

*University Paris 7 - CNRS, Department of Mathematics, Case 7012, 2 Place Jussieu, 75251 Paris Cedex 05, France. E-mail: elibou@logique.jussieu.fr

2. Different notions of dimensions can be attached to definable sets (or infinite intersections of definable sets, which we call *infinitely definable, or ∞-definable, sets*). In the case of algebraically closed fields, all such possible notions of abstract dimension must coincide and be equal to the classical algebraic dimension. In the other cases, these dimensions may be different, some may take infinite ordinal values or may be defined only for some special classes of definable (or ∞-definable) sets.

3. If K is any of the above mentioned fields, and if H is an algebraic group defined over K, then the group $H(K)$ of the K-rational points of H is a definable group. But there are "new" families of definable groups which are not of this form.

In fact, it is precisely the study of certain specific families of "new" definable groups of finite dimension which are at the center of the applications to Diophantine geometry. We will not attempt here to explain how the model theoretic analysis of the finite rank definable groups yields these applications. There have been in recent years many surveys and presentations of the subject to which we refer the reader (see for example, [4],[5], [14], [22] or [28]). We will come back to this subject, but very briefly, at the end in Section 3.5..

The first general question raised by the existence of these new definable groups is that of their relation to the classical algebraic groups. Remark that this question already makes sense in the context of "pure" algebraically closed fields, about the class of definable (= constructible) groups. In that case, it is true that any constructible group in an algebraically closed field K is constructibly isomorphic to the K-rational points of an algebraic group defined over K (see for example [3] or [23]).

Let us now consider briefly the case of a field K of characteristic $p > 0$ which is separably closed and not perfect. Then the class of constructible sets is no longer closed under projection and there are many definable groups which are not constructible, the most obvious one being K^p. There are also some groups which are proper intersections of infinite descending chains of definable groups: for example, $K^{p^\infty} (= \bigcap_n K^{p^n})$, the field of infinitely p-divisible elements of the multiplicative group, or $\bigcap_n p^n A(K)$, for A an Abelian variety defined over K.

It is nevertheless true, as we will see, that every definable group in K is definably isomorphic to the K-rational points of an algebraic group defined over K. Furthermore, as in the classical case of one-dimensional algebraic groups, it is possible to give a complete description, up to definable isomorphism, of the one-dimensional infinitely definable groups.

There are results of similar type for the other classes of enriched fields mentioned above. In this short paper, we will concentrate mainly on the case of separably closed fields (in Section 3.). Before this, in Section 2., we will only very briefly present the model theoretic setting for two other examples of "enriched" fields, in characteristic zero, differentially closed fields and generic difference fields. We hope this will give the reader an idea of what the common features and the differences might be in the model theoretic analysis of these different classes of fields.

Finally, there are of course many other classes of fields whose model theory has been extensively developed in the past years with many connections to algebra, semi-algebraic or subanalytic geometry, and which we are not going to mention here: for example, valued fields, ordered fields, "o-minimal" expansions of the real

field...

2. Two short examples

We will just very briefly describe the two characteristic zero examples mentioned above.

2.1. Differentially closed fields of characteristic zero

We consider a field K of characteristic zero, with a *derivation* δ, that is, an additive map from K to K which satisfies that for all x, y in K, $\delta(xy) = x\delta(y) + y\delta(y)$. We define the ring $K_\delta[X]$ of *differential polynomials* over K to be the ring of polynomials in infinitely many variables $K[X, \delta(X), \delta^2(X), \cdots, \delta^n(X), \cdots]$. The *order* of the differential polynomial $f(X)$ in $K_\delta[X]$ is -1 if $f \in K$ and otherwise the largest n such that $\delta^n(X)$ occurs in $f(X)$ with non zero coefficient. We say that K is *differentially closed* if for any non-constant differential polynomials $f(X)$ and $g(X)$, where the order of g is strictly less than the order of f, there is a z such that $f(z) = 0$ and $g(z) \neq 0$. In model theoretic terms, this means exactly that K is existentially closed.

From now on we suppose that (K, δ) is a large differentially closed field (a universal domain).

We say that $F \subseteq K^n$ is a *δ-closed set*, if there are $f_1, \cdots, f_r \in K_\delta[X_1, \cdots, X_n]$ such that $F = \{(a_1, \cdots, a_n) \in K^n; f_1(a_1, \cdots, a_n) = \cdots = f_r(a_1, \cdots, a_n) = 0\}$. The ring $K_\delta[X_1, \cdots, X_n]$ is of course not Noetherian but the δ-closed sets (which correspond to radical differential ideals) form the closed sets of a Noetherian topology on K, the *δ-topology*.

We now consider the *δ-constructible* sets, that is, the finite Boolean combinations of δ-closed sets. This class is closed under projection (this is quantifier elimination for the theory), hence the definable sets (we call them *δ-definable* sets) are exactly the δ-constructible sets. To every δ-definable set one can associate a dimension (the Morley rank) which can take infinite countable ordinal values.

There are "new" definable groups, which are not of the form $H(K)$ for any algebraic group H. In particular, any $H(K)$ will have infinite dimension. In contrast, the *field of constants of* K, $Cons(K) = \{a \in K; \delta(x) = 0\}$, is a δ-closed set which is not constructible; it is an algebraically closed subfield of K and has dimension one.

Nevertheless the following is true:

Proposition 1 ([21]) *Let G be a δ-definable group in K. Then there is an algebraic group H, defined over K, such that G is definably isomorphic to a (δ-definable) subgroup of $H(L)$.*

For the many more existing results on δ-definable groups, we refer the reader to [20], or from the differential algebra point of view, to [8].

2.2. Generic difference fields

We now consider an algebraically closed field K with an automorphism σ. We say that (K, σ) is a generic difference field if every difference equation which has a solution in an extension of K has a solution in K. The theory of generic difference fields has been extensively studied in [9] and [10].

Let us suppose that (K, σ) is a generic difference field in characteristic zero. We consider the ring of *σ-polynomials*,

$$K_\sigma[X_1, \cdots, X_n] = K[X_1, \cdots, X_n, \sigma(X_1), : \cdots, \sigma(X_n), \sigma^2(X_1), \cdots, \sigma^2(X_n), \cdots].$$

We say that $F \subseteq K^n$ is a *$f\sigma$-closed set* if there are $f_1, \cdots, f_r \in K_\sigma[X_1, \cdots, X_n]$ such that $F = \{(a_1, \cdots, a_n) \in K^n : f_1(a_1, \cdots, a_n) = \cdots = f_r(a_1, \cdots, a_n) = 0\}$. The σ-closed sets form the closed sets of a Noetherian topology on K, the *σ-topology.* The class of σ-definable sets is the closure under finite Boolean operations and projections of the σ-closed sets.

Again there are "new" σ-definable groups. For example, the field $Fix(K) = \{a \in K : \sigma(a) = a\}$, the fixed field of σ in K, is a σ-closed set of dimension one.

Here the best result possible for arbitrary σ-definable groups is the following:

Proposition 2 ([18]) *Let G be a group definable in (K, σ). Then there are an algebraic group H defined over K, a finite normal subgroup N_1 of G, a σ-definable subgroup H_1 of $H(K)$ and a finite normal subgroup N_2 of H_1, such that G/N_1 and H_1/N_2 are σ-definably isomorphic.*

The analysis of groups of finite dimension is one of the main tools in Hrushovski's proof of the Manin-Mumford conjecture in [15].

3. Separably closed fields of finite degree of imperfection

Separably closed fields are particularly interesting from the model theoretic point of view for many reasons, in addition to the fact that they form the framework for Hrushovski's proof of the Mordell-Lang conjecture in charactersitic p. Let us just mention one reason here: they are the only fields known to be stable and non superstable, and in fact it is conjectured that they are the only existing ones.

We will just focus on the main properties of the groups that are definable in a separably closed field of finite degree of imperfection, but we need first to introduce some notation and recall some basic facts (see [11]).

3.1. Some basic facts and notation

Let L be a separably closed field of charcteristic $p > 0$ and of finite degree of imperfection which is not perfect, i.e., L has no proper separable algebraic extension, and $|L : L^p| = p^\nu$, with $0 < \nu$. In order to avoid confusion we denote the Cartesian product of k copies of L by $L^{\times k}$.

A subset $B = \{b_1, \cdots, b_\nu\}$ of L is called a *p-basis* of L if the *set of p-monomials of B*, $\{M_j := \prod_{i=1}^{\nu} b_i^{j(i)}; j \in p^\nu\}$ forms a linear basis of L over L^p. Each element x

in L can be written in a unique way as $x = \sum_{j \in p^\nu} x_j{}^p M_j$. **From now on we fix a p-basis B of L and the M_j's, with $j \in p^\nu$, always denote the p-monomials of B.** We suppose that L is large (a universal domain, or in model theoretic terms, saturated) and we fix some small separably closed subfield K of L, containing B and of same degree of imperfection ν.

We let f_j denote the map which to x associates x_j. The x_j's are called the *p-components of x of level one.* More generally, one can associate to x a tree of countable height indexed by $(p^\nu)^{<\omega}$, which we call the *tree of p-components of x.* For $\sigma \in (p^\nu)^{<\omega}$, we define x_σ by induction: $x_\emptyset = x$ and if $\tau \in (p^\nu)^n$, and $j \in p^\nu$, we let $x_{(\tau,j)}$ be equal to $f_j(x_\tau)$; $x_{(\tau,j)}$ is called a p-component of x of level $n+1$.

We will also use the notation $a_\infty := (a_\sigma)_{\sigma \in (p^\nu)^{<\omega}}$, for $a \in L$.

The ring $K[X_\infty]$. $K[X_\infty]$ is the polynomial ring in countably many indeterminates indexed in a way which will allow the natural substitution by the p-components of elements: for X a single variable, $X_\infty := (X_\sigma)_{\sigma \in (p^\nu)^{<\omega}}$, and for $X = (Y_1, \ldots, Y_k)$ a k-tuple of variables, $X_\infty := ((Y_1)_\infty, \ldots, (Y_k)_\infty)$. The ring $K[X_\infty]$ is a countable union of Noetherian rings, hence each ideal is countably generated. We let $I^0(X)$ denote the ideal of $K[X_\infty]$ generated by the polynomials $X_\sigma - \sum_{j \in p^\nu} X^p_{(\sigma,j)} M_j$, $\sigma \in (p^\nu)^{<\omega}$.

3.2. The λ-topology

Given a set of polynomials S of $K[X_\infty]$, let $V(S) = \{a \in L^{\times k} : f(a_\infty) = 0 \text{ for all } f \in S\}$. Such a $V(S)$ is called *λ-closed* (with parameters in K or over K) in L.

Given $A \subseteq L^{\times k}$, we define its *canonical ideal* $I(A)$ over K, $I(A) := \{f \in K[X_\infty] : f(a_\infty) = 0 \text{ for all } a \in A\}$.

The λ-closed subsets of $L^{\times k}$ form the closed sets of the *λ-topology* on $L^{\times k}$. This topology is not Noetherian but is the limit of countably many Noetherian topologies.

Let C be a commutative K-algebra. An ideal I of C is *separable* if, for all $c_j \in C$, $j \in p^\nu$, if $\sum_{j \in p^\nu} c_j^p M_j \in I$, then each $c_j \in I$.

Fact 3 ("Nullstellensatz") *1. The map $A \mapsto I(A)$ induces a bijection between λ-closed subsets of the affine space L^k which are defined over K and ideals of $K[X_\infty]$ which are separable and contain $I^0(X)$. The inverse map is $I \mapsto V(I)$.*

Now for the basic properties of the first-order theory:

Fact 4 *1. The theory of separably closed fields of characteristic p, of degree of imperfection ν, and with p-basis $\{b_1, \ldots, b_\nu\}$ is complete and admits elimination of quantifiers and elimination of imaginaries in the language*

$$\mathcal{L}_{p,\nu} = \{0, 1, +, -, .\} \cup \{b_1, \ldots, b_\nu\} \cup \{f_i; i \in p^\nu\}.$$

In particular, any definable set is λ-constructible, that is, a finite Boolean combinations of definable λ-closed sets. Remark that it is impossible to associate to an arbitrary definable set a well-behaved notion of dimension: indeed, such a

dimension would need to be invariant under definable bijections, but for every n the map λ_n, defined by $\lambda_n(x) := (x_\sigma)_{\sigma \in (p^\nu)^n}$, is a definable bijection between L and $L^{\times p^{\nu n}}$. But some ∞-definable sets will have a well-defined dimension, for example the field $L^{p^\infty} := \bigcap_n L^{p^n}$, which is the biggest algebraically closed subfield of L, has dimension one. In fact, L^{p^∞} is the unique (up to definable isomorphism) infinitely definable field of dimension one ([19], [13]).

3.3. Definable groups

Again, amongst the definable groups, one finds the "classical" ones, that is groups of the form $H(L)$ for H any algebraic group defined over L. These groups have certain specific properties which are not true of all the definable groups in L. Recall that a definable subset X of G is said to be *generic* if G is covered by a finite number of translates of X, and an element of G is *generic for the group* if every definable set which contains it is generic. In an algebraic group, generics in the topological sense coincide with generics for the algebraic group. Recall also that a definable group is said to be *connected* if it has no proper definable subgroup of finite index, and *connected-by-finite* if it has a definable connected subgroup of finite index.

Proposition 5 ([6], [13]) *Let H be an algebraic group defined over K. Then $H(L)$ is connected-by-finite. If H is connected (hence irreducible as an algebraic group), then $H(L)$ is connected (and irreducible for the λ-topology) and if $a \in H(L)$ is a generic point, then the ideal $I(a) = \{f \in K[X_\infty] : f(a_\infty) = 0\}$ is minimal amongst the ideals $I(h)$, for $h \in H(L)$.*

The above says that in the group $H(L)$, the generics in the topological sense coincide with generics for the group. In an arbitrary group defined in L, this need not be the case.

Consider the definable bijection f from L to L defined in the following way: if $x \in L \setminus L^p$, $f(x) = x^p$, if $x \in L^p \setminus L^{p^2}$, $f(x) = x^{1/p}$, if $x \in L^{p^2}$, $f(x) = x$.

Transporting addition through f, one gets a group on L again, $G := (L, *)$, definably isomorphic to $(L, +)$, hence connected. The set L itself is of course λ-closed and irreducible with associated ideal $I(L) = I^0(X)$. The ideal associated to the (group) generic of $(L, *)$ is generated by $I^0(X)$ and $\{X_i = 0 : i \in p^\nu, i \neq 0\}$, and strictly contains $I^0(X)$.

This question of the uniqueness of the notion of generic is not the only one posing problems for arbitrary definable groups in L. For example, there is no reason, coming from general properties of stable (non superstable) theories, which a priori forces all these definable groups to be connected-by-finite.

Nevertheless, one can in fact show that the situation is as close to the classical one as it could be:

Proposition 6 [6] *Every definable group G in L is connected-by-finite and is definably isomorphic to the group of L-rational points of an algebraic group H defined over L.*

One more remark, in the case of algebraic groups, by Prop. 5, irreducibility transfers down to the set of L-rational points. But this is not the case for an arbitrary variety: if one considers for example the irreducible variety defined by the equation $Y^{p^m}X + Z^{p^m} = 0$, for $m \geq 1$, then the λ-closed set $V(L)$ is no longer irreducible in the sense of the λ-topology.

3.4. Minimal groups

The previous result enables us to give a complete description of groups of dimension one, and more generally of some classes of commutative groups.

We say that an ∞-definable set D is *minimal* if any definable subset of G is finite or co-finite. If D is actually definable, then we say that D is *strongly minimal.*

The minimal groups are exactly the connected groups of dimension (U-rank) equal to one. A minimal group must be commutative.

From the basic properties of commutative algebraic groups over an algebraically closed field of characteristic p and Proposition 6, one can deduce:

Lemma 7 *Let G be a minimal group ∞-definable in L, then G has exponent p or G is divisible.*

We first consider the commutative groups of exponent p:

Proposition 8 [7] *Let G be a commutative ∞-definable group of exponent p definable in L. Then G is definably isomorphic to a λ-closed subgroup of the additive group $(L,+)$. Furthermore, if G is definable, then it is definably isogenous to the group of L-rational points of a vector group.*

Note that even when G is connected it is not necessarily definably isomorphic to the group of rational points of a vector group.

Then we consider the commutative divisible groups, which we show to be exactly the ones that were considered by Hrushovski in [13]:

Proposition 9 [7] *1. Let G be any ∞-definable commutative divisible group in L. Then G is definably isomorphic to some $p^\infty A(L) := \bigcap_n p^n A(L)$, for A a semi-Abelian variety defined over L.*

2. If A is a semi-Abelian variety defined over L, $p^\infty A(L)$, which is the maximal divisible subgroup of $A(L)$ is also the smallest ∞-definable subgroup of $A(L)$ which is Zariski dense in A.

Finally, this analysis, together with some results from [11] and [13], yields the full description of minimal groups.

Before stating the actual result, let us give some last definitions. The group G is said to be *of linear type* if for every n, every definable subgroup of $G^{\times n}$ is a finite Boolean combination of translates of definable subgroups of $G^{\times n}$. We define the *transcendence rank over K* of a group G, defined over K, to be the maximum of $\{\text{tr.degree}(K(g_\infty), K) : g \in G\}$.

Proposition 10 *Let G be an ∞-definable minimal group in L.*

1. *Either G is not of linear type and then,*
   - *G is definably isomorphic to the multiplicative group $((L^{p^\infty})^*, \cdot)$,*
   - *or G is definably isomorphic to $E(L^{p^\infty})$ for E an elliptic curve defined over L^{p^∞},*
   - *or G is definably isogenous to $(L^{p^\infty}, +)$. (isogenous here cannot be replaced by isomorphic).*
2. *Or G is of linear type and then,*
   - *G is divisible and G is definably isomorphic to $p^\infty A(L)$ for some simple Abelian variety A defined over K which is not isogenous to an Abelian variety defined over L^{p^∞},*
   - *or G is of exponent p and is definably isomorphic to a minimal λ-closed subgroup of $(L, +)$.*

   In the divisible case G has finite transcendence rank; in the exponent p case, all transcendence ranks are possible.

The induced module-type structure on the minimal groups of exponent p and of linear type is analyzed in [2].

A short word about some of the tools involved in the proofs of Propositions 6 and 10: the proofs of 6, 1 and 2 all involve at some point the classical theorem of Weil's constructing an algebraic group from a generic group law on a variety, or some generalizations of this theorem to an abstract model theoretic context. In the specific case of separably closed fields, another fundamental tool is the analysis of the properties of the *Λ_n-functors*, naturally associated to the maps λ_n: for each n, Λ_n is a covariant functor from the category of varieties V defined over K to itself, with the property that the L-rational points of the variety $\Lambda_n V$ are exactly the image by the map λ_n of the L-rational points of V. In the case of an algebraic group defined over K, Λ_1 is equal to the composition of the inverse of the Frobenius and of the classical Weil restriction of scalars functor from $K^{1/p}$ to K.

Finally, the way we have stated Proposition 10 uses the fact that if a minimal group is not of linear type, then it is non orthogonal to L^{p^∞} (and hence definably isogenous to the L^{p^∞}-rational points of some definable group over L^{p^∞}). The only known proof of this so far uses the powerful abstract machinery of Zariski structures from [16]. This dichotomy result, for the particular case of groups of the form $p^\infty A(L)$, is essential in Hrushovski's proof of the Mordell-Lang conjecture in characteristic p, which is still the only existing proof for the general case. In a recent paper Pillay and Ziegler ([24]), show that, with some extra assumptions on A, one can replace in this proof the heavy Zariski structure argument by a much more elementary one. These extra assumptions are satisfied when A is an *ordinary* semi-Abelian variety (i.e. A has the maximum possible number of p^n-torsion points for every n), case which was already covered by previous non model-theoretic proofs (see [1]).

3.5. Final remarks and questions

As we have already mentioned earlier, the groups of finite dimension definable in these "enriched" theories of fields play a major role in the applications of

model theory to Diophantine geometry. In the characteristic zero case, the relevant groups are the definable subgroups of the group of rational points of Abelian varieties in differentially closed fields (Mordell-Lang conjecture for function fields [13]), in generic difference fields (the Manin-Mumford conjecture [15], [5] and the Tate-Voloch conjecture for semi-Abelian varieties defined over $\mathbb{Q}_p$ [25], [26]). In the characteristic p case, the relevant groups are: the ∞-definable divisible subgroups of the group of rational points of semi-Abelian varieties in separably closed fields (the Mordell-Lang conjecture for function fields [13]) and the definable subgroups of the additive groups in generic difference fields of characteristic p (Drinfeld modules [27]).

One should note that, in fact, separably closed fields are just another instance of a *field with extra operators* (derivations or automorphisms): one can equip any separably closed field L of finite degree of imperfection, with an infinite family of Hasse derivations in such a way that the resulting structure is bi-definably equivalent with L considered as a structure in the language described in section 3.2.. There are many interesting other possible types of "enriched" fields in this sense where the complete analysis of the model theoretic structure remains to be done.

Finally, one crucial step towards possible further applications of the fine study of finite rank definable sets to geometry would be an understanding of the structure induced on the so-called *trivial* or *disintegrated* definable (or infinitely definable) minimal sets, that is the minimal sets such that the induced pregeometry is disintegrated. This condition immediately rules out definable groups. The absence of any well-understood algebraic structure living on these "trivial" sets makes them very difficult to analyze. The only results obtained so far are in the context of differentially closed fields of characteristic 0: Hrushovski ([12]), building on some results of Jouanolou ([17]), showed that in any trivial strongly minimal set defined by a differential equation of order one, the induced pregeometry is locally finite. The question of whether this is true for higher order equations is still open.

References

[1] D. Abramovic & F. Voloch, Towards a proof of the Mordell-Lang conjecture in characteristic p, *Intern. Math. Research Notices (IMRN)*, 2 (1992), 103-115.

[2] T. Blossier, Ensembles minimaux localement modulaires, Thèse de Doctorat, Université Paris 7, 2001.

[3] E. Bouscaren, Model-theoretic versions of Weil's theorem on pre-groups, in *The Model Theory of Groups*, (A. Nesin & A. Pillay, editors), Notre Dame University Press, 1989.

[4] E. Bouscaren, Proof of the Mordell-Lang conjecture for function fields, in *Model theory and algebraic geometry* (E. Bouscaren, editor), Lecture Notes in Mathematics, Vol. 1696, Springer-Verlag, 1998.

[5] E. Bouscaren, Théorie des modèles et conjecture de Manin-Mumford (d'après Ehud Hrushovski), *Séminaire Bourbaki*, Vol. 1999/2000, Astérisque No. 276 (2002), 137–159.

[6] E. Bouscaren & Françoise Delon, Groups definable in separably closed fields, *Transactions of the A.M.S.*, 354 (2002), 945–966.

[7] E. Bouscaren & Françoise Delon, Minimal groups in separably closed fields, *The Journal of Symbolic Logic*, 67 (2002), 239–259.
[8] A. Buium, *Differential Algebra and Diophantine Geom.*, Hermann, Paris, 1994.
[9] Z. Chatzidakis & E. Hrushovski, The model theory of difference fields, *Transactions of the A.M.S*, Vol. 351 (1999), 2997-3071.
[10] Z. Chatzidakis, E. Hrushovski & Y. Peterzil, The model theory of difference fields II, *Proceedings of the London Math. Soc.* (to appear).
[11] F. Delon, Separably closed fields, in *Model Theory and Algebraic Geometry*, E. Bouscaren (Ed.), Lecture Notes in Mathematics 1696, Springer-Verlag, 1998.
[12] E. Hrushovski, ODE's of order 1 and a generalisation of a theorem of Jouanolou's, Manuscript, 1995.
[13] E. Hrushovski, The Mordell-Lang conjecture for function fields, *Journal of the A.M.S.*, 9 (1996), 667–690.
[14] E. Hrushovski, Geometric model theory, in *Proceedings of the International Congress of Mathematicians*, Berlin, Vol. I (1998), Doc. Math., 281–302.
[15] E. Hrushovski, The Manin-Mumford conjecture and the model theory of difference fields,*Annals of Pure and Applied Logic*, 112 (2001), 43–115.
[16] E. Hrushovski & B. Zilber, Zariski Geometries, *Journal of the A.M.S.*, 9 (1996), 1-56.
[17] J.P. Jouanolou, Hypersurfaces solutions d'une équation de Pfaff analytique, *Mathematische Annalen*, 232 (1978), 239–245.
[18] P. Kowalski & A. Pillay, A note on groups definable in difference fields, preprint, 2000.
[19] M. Messmer, Groups and fields interpretable in separably closed fields, *Transactions of the A.M.S.*, 344 (1994), 361–377.
[20] A. Pillay, Differential algebraic groups and the number of countable differentially closed fields, in *Model Theory of Fields*, D. Marker, M. Messmer & A. Pillay, Lecture Notes in Logic 5, Springer, 1996.
[21] A. Pillay, Some foundational questions concerning differential algebraic groups, *Pacific Journal of Math.*, 179 (1997), 179-200.
[22] A. Pillay, Model Theory and Diophantine geometry, *Bulletin of the A.M.S.*, 34 (1997), 405–422.
[23] A. Pillay, *Model theory of algebraically closed fields*, in *Model theory and algebraic geometry* (E. Bouscaren, editor), Lecture Notes in Mathematics, Vol. 1696, Springer-Verlag, 1998.
[24] A. Pillay & M. Ziegler, Jet spaces of varieties over differential and difference fields, preprint, 2002.
[25] T. Scanlon, p-adic distance from torsion points of semi-Abelian varieties, *Journal für dir Reine und Angewandte Mathematik*, 499 (1998), 225-236.
[26] T. Scanlon, The conjecture of Tate & Voloch on p-adic proximity to torsion, *Intern. Math. Research Notices (IMRN)*, 17 (1999), 909-914.
[27] T. Scanlon, Diophantine geometry of the torsion of a Drinfeld module, preprint 1999.
[28] T. Scanlon, Diophantine geometry from model theory, *Bulletin of Symbolic Logic*, 7 (2001), 37–57.

ICM 2002 · Vol. II · 13–23

Motivic Integration and the Grothendieck Group of Pseudo-Finite Fields

J. Denef* F. Loeser†

Abstract

Motivic integration is a powerful technique to prove that certain quantities associated to algebraic varieties are birational invariants or are independent of a chosen resolution of singularities. We survey our recent work on an extension of the theory of motivic integration, called arithmetic motivic integration. We developed this theory to understand how p-adic integrals of a very general type depend on p. Quantifier elimination plays a key role.

2000 Mathematics Subject Classification: 03C10, 03C98, 12E30, 12L12, 14G15, 14G20, 11G25, 11S40, 12L10, 14F20.
Keywords and Phrases: Motivic integration, p-adic integration, Quantifier elimination.

1. Introduction

Motivic integration was first introduced by Kontsevich [20] and further developed by Batyrev [3][4], and Denef-Loeser [8][9][12]. It is a powerful technique to prove that certain quantities associated to algebraic varieties are birational invariants or are independent of a chosen resolution of singularities. For example, Kontsevich used it to prove that the Hodge numbers of birationally equivalent projective Calabi-Yau manifolds are equal. Batyrev [3] obtained his string-theoretic Hodge numbers for canonical Gorenstein singularities by motivic integration. These are the right quantities to establish several mirror-symmetry identities for Calabi-Yau varieties. For more applications and references we refer to the survey papers [11] and [21]. Since than, several other applications to singularity theory were discovered, see e.g. Mustaţă [24].

In the present paper, we survey our recent work [10] on an extension of the theory of motivic integration, called arithmetic motivic integration. We developed

*Department of Mathematics, University of Leuven, Celestijnenlaan 200 B, 3001 Leuven, Belgium. E-mail: Jan.Denef@wis.kuleuven.ac.be

†Département de Mathématiques et Applications, École Normale Supérieure, 45 rue d'Ulm, 75230 Paris Cedex 05, France (UMR 8553 du CNRS). E-mail: Francois.Loeser@ens.fr

this theory to understand how p-adic integrals of a very general type depend on p. This is used in recent work of Hales [18] on orbital integrals related to the Langlands program. Arithmetic motivic integration is tightly linked to the theory of quantifier elimination, a subject belonging to mathematical logic. The roots of this subject go back to Tarski's theorem on projections of semi-algebraic sets and to the work of Ax-Kochen-Ersov and Macintyre on quantifier elimination for Henselian valued fields (cf. section 4). We will illustrate arithmetic motivic integration starting with the following concrete application. Let X be an algebraic variety given by equations with integer coefficients. Denote by $N_{p,n}$ the cardinality of the image of the projection $X(\mathbf{Z}_p) \to X(\mathbf{Z}/p^{n+1})$, where $\mathbf{Z}_p$ denotes the p-adic integers. A conjecture of Serre and Oesterlé states that $P_p(T) := \sum_n N_{p,n} T^n$ is rational. This was proved in 1983 by Denef [7] using quantifier elimination, expressing $P_p(T)$ as a p-adic integral over a domain defined by a formula involving quantifiers. This gave no information yet on how $P_p(T)$ depends on p. But recently, using arithmetic motivic integration, we proved:

Theorem 1.1. *There exists a canonically defined rational power series $P(T)$ over the ring $\mathrm{K}_0^{mot}(\mathrm{Var}_{\mathbf{Q}}) \otimes \mathbf{Q}$, such that, for $p \gg 0$, $P_p(T)$ is obtained from $P(T)$ by applying to each coefficient of $P(T)$ the operator N_p.*

Here $\mathrm{K}_0(\mathrm{Var}_{\mathbf{Q}})$ denotes the Grothendieck ring of algebraic varieties over $\mathbf{Q}$, and $\mathrm{K}_0^{mot}(\mathrm{Var}_{\mathbf{Q}})$ is the quotient of this ring obtained by identifying two varieties if they have the same class in the Grothendieck group of Chow motives (this is explained in the next section). Moreover the operator N_p is induced by associating to a variety over $\mathbf{Q}$ its number of rational points over the field with p elements, for $p \gg 0$.

As explained in section 8 below, this theorem is a special case of a much more general theorem on p-adic integrals. There we will also see how to canonically associate a "virtual motive" to quite general p-adic integrals. A first step in the proof of the above theorem is the construction of a canonical morphism from the Grothendieck ring $\mathrm{K}_0(\mathrm{PFF}_{\mathbf{Q}})$ of the theory of pseudo-finite fields of characteristic zero, to $\mathrm{K}_0^{mot}(\mathrm{Var}_{\mathbf{Q}}) \otimes \mathbf{Q}$. Pseudo-finite fields play a key role in the work of Ax [1] that leads to quantifier elimination for finite fields [19][14][5]. The existence of this map is interesting in itself, because any generalized Euler characteristic, such as the topological Euler characteristic or the Hodge-Deligne polynomial, can be evaluated on any element of $\mathrm{K}_0^{mot}(\mathrm{Var}_{\mathbf{Q}}) \otimes \mathbf{Q}$, and hence also on any logical formula in the language of fields (possibly involving quantifiers). All this will be explained in section 2. In section 3 we state Theorem 3.1, which is a stronger version of Theorem 1.1 that determines $P(T)$. A proof of Theorem 3.1 is outlined in section 7, after giving a survey on arithmetic motivic integration in section 6.

2. The Grothendieck group of pseudo-finite fields

Let k be a field of characteristic zero. We denote by $\mathrm{K}_0(\mathrm{Var}_k)$ the Grothendieck ring of algebraic varieties over k. This is the group generated by symbols $[V]$ with V an algebraic variety over k, subject to the relations $[V_1] = [V_2]$ if V_1 is isomorphic to V_2, and $[V \setminus W] = [V] - [W]$ if W is a Zariski closed subvariety of W. The ring

multiplication on $\mathrm{K}_0(\mathrm{Var}_k)$ is induced by the cartesian product of varieties. Let $\mathbf{L}$ be the class of the affine line over k in $\mathrm{K}_0(\mathrm{Var}_k)$. When V is an algebraic variety over $\mathbf{Q}$, and p a prime number, we denote by $N_p(V)$ the number of rational points over the field $\mathbf{F}_p$ with p elements on a model $\tilde{V}$ of V over $\mathbf{Z}$. This depends on the choice of a model $\tilde{V}$, but two different models will yield the same value of $N_p(V)$, when p is large enough. This will not cause any abuse later on. For us, an algebraic variety over k does not need to be irreducible; we mean by it a reduced separated scheme of finite over k.

To any projective nonsingular variety over k one associates its Chow motive over k (see [27]). This is a purely algebro-geometric construction, which is made in such a way that any two projective nonsingular varieties, V_1 and V_2, with isomorphic associated Chow motives, have the same cohomology for each of the known cohomology theories (with coefficients in a field of characteristic zero). In particular, when k is $\mathbf{Q}$, $N_p(V_1) = N_p(V_2)$, for $p \gg 0$. For example two elliptic curves define the same Chow motive iff there is a surjective morphism from one to the other. We denote by $\mathrm{K}_0^{mot}(\mathrm{Var}_k)$ the quotient of the ring $\mathrm{K}_0(\mathrm{Var}_k)$ obtained by identifying any two nonsingular projective varieties over k with equal associated Chow motives. From work of Gillet and Soulé [15], and Guillén and Navarro Aznar[17], it directly follows that there is a unique ring monomorphism from $\mathrm{K}_0^{mot}(\mathrm{Var}_k)$ to the Grothendieck ring of the category of Chow motives over k, that maps the class of a projective nonsingular variety to the class of its associated Chow motive. What is important for the applications, is that any generalized Euler characteristic, which can be defined in terms of cohomology (with coefficients in a field of characteristic zero), factors through $\mathrm{K}_0^{mot}(\mathrm{Var}_k)$. With a generalized Euler characteristic we mean any ring morphism from $\mathrm{K}_0(\mathrm{Var}_k)$, for example the topological Euler characteristic and the Hodge-Deligne polynomial when $k = \mathbf{C}$. For $[V]$ in $\mathrm{K}_0^{mot}(\mathrm{Var}_k)$, with $k = \mathbf{Q}$, we put $N_p([V]) = N_p(V)$; here again this depends on choices, but two different choices yield the same value for $N_p([V])$, when p is large enough.

With a ring formula φ over k we mean a logical formula build from polynomial equations over k, by taking Boolean combinations and using existential and universal quantifiers. For example, $(\exists x)(x^2 + x + y = 0$ and $4y \neq 1)$ is a ring formula over $\mathbf{Q}$. The mean purpose of the present section is to associate in a canonical way to each such formula φ an element $\chi_c([\varphi])$ of $\mathrm{K}_0^{mot}(\mathrm{Var}_k) \otimes \mathbf{Q}$. One of the required properties of this association is the following, when $k = \mathbf{Q}$: If the formulas φ_1 and φ_2 are equivalent when interpreted in $\mathbf{F}_p$, for all large enough primes p, then $\chi_c([\varphi_1])$ $= \chi_c([\varphi_2])$. The natural generalization of this requirement, to arbitrary fields k of characteristic zero, is the following: If the formulas φ_1 and φ_2 are equivalent when interpreted in K, for all pseudo-finite fields K containing k, then $\chi_c([\varphi_1]) = \chi_c([\varphi_2])$. We recall that a pseudo-finite field is an infinite perfect field that has exactly one field extension of any given finite degree, and over which each absolutely irreducible variety has a rational point. For example, infinite ultraproducts of finite fields are pseudo-finite. J. Ax [1] proved that two ring formulas over $\mathbf{Q}$ are equivalent when interpreted in $\mathbf{F}_p$, for all large enough primes p, if and only if they are equivalent when interpreted in K, for all pseudo-finite fields K containing $\mathbf{Q}$. This shows that the two above mentioned requirements are equivalent when $k = \mathbf{Q}$. In fact, we

will require much more, namely that the association $\varphi \longmapsto \chi_c([\varphi])$ factors through the Grothendieck ring $\mathrm{K}_0(\mathrm{PFF}_k)$ of the theory of pseudo-finite fields containing k. This ring is the group generated by symbols $[\varphi]$, where φ is any ring formula over k, subject to the relations $[\varphi_1 \text{ or } \varphi_2] = [\varphi_1] + [\varphi_2] - [\varphi_1 \text{ and } \varphi_2]$, whenever φ_1 and φ_2 have the same free variables, and the relations $[\varphi_1] = [\varphi_2]$, whenever there exists a ring formula ψ over k that, when interpreted in any pseudo-finite field K containing k, yields the graph of a bijection between the tuples of elements of K satisfying φ_1 and those satisfying φ_2. The ring multiplication on $\mathrm{K}_0(\mathrm{PFF}_k)$ is induced by the conjunction of formulas in disjoint sets of variables. We can now state the following variant of a theorem of Denef and Loeser [10].

Theorem 2.1. *There exists a unique ring morphism*

$$\chi_c : \mathrm{K}_0(\mathrm{PFF}_k) \longrightarrow \mathrm{K}_0^{mot}(\mathrm{Var}_k) \otimes \mathbf{Q}$$

satisfying the following two properties:
(i) For any formula φ which is a conjunction of polynomial equations over k, the element $\chi_c([\varphi])$ equals the class in $\mathrm{K}_0^{mot}(\mathrm{Var}_k) \otimes \mathbf{Q}$ of the variety defined by φ.
(ii) Let X be a normal affine irreducible variety over k, Y an unramified Galois cover [1] of X, and C a cyclic subgroup of the Galois group G of Y over X. For such data we denote by $\varphi_{Y,X,C}$ a ring formula, whose interpretation in any field K containing k, is the set of K-rational points on X that lift to a geometric point on Y with decomposition group C (i.e. the set of points on X that lift to a K-rational point of Y/C, but not to any K-rational point of Y/C' with C' a proper subgroup of C). Then

$$\chi_c([\varphi_{Y,X,C}]) = \frac{|C|}{|\mathrm{N_G}(C)|}\chi_c([\varphi_{Y,Y/C,C}]),$$

where $\mathrm{N_G}(C)$ is the normalizer of C in G.

Moreover, when $k = \mathbf{Q}$, we have for all large enough primes p that $N_p(\chi_c([\varphi]))$ equals the number of tuples in $\mathbf{F}_p$ that satisfy the interpretation of φ in $\mathbf{F}_p$.

The proof of the uniqueness goes as follows: From quantifier elimination for pseudo-finite fields (in terms of Galois stratifications, cf. the work of Fried and Sacerdote [14][13, §26]), it follows that every ring formula over k is equivalent (in all pseudo-finite fields containing k) to a Boolean combination of formulas of the form $\varphi_{Y,X,C}$. Thus by (ii) we only have to determine $\chi_c([\varphi_{Y,Y/C,C}])$, with C a cyclic group. But this follows directly from the following recursion formula:

$$|C|\,[Y/C] = \sum_{A \text{ subgroup of } C} |A|\chi_c([\varphi_{Y,Y/A,A}]).$$

This recursion formula is a direct consequence of (i), (ii), and the fact that the formulas $\varphi_{Y,Y/C,A}$ yield a partition of Y/C. The proof of the existence of the morphism χ_c is based on the following. In [2], del Baño Rollin and Navarro Aznar associate to any representation over $\mathbf{Q}$ of a finite group G acting freely on an affine variety Y over k, an element in the Grothendieck group of Chow motives over k. By

[1] Meaning that Y is an integral étale scheme over X with $Y/G \cong X$, where G is the group of all endomorphisms of Y over X.

linearity, we can hence associate to any $\mathbf{Q}$-central function α on G (i.e. a $\mathbf{Q}$-linear combination of characters of representations of G over $\mathbf{Q}$), an element $\chi_c(Y,\alpha)$ of that Grothendieck group tensored with $\mathbf{Q}$. Using Emil Artin's Theorem, that any $\mathbf{Q}$-central function α on G is a $\mathbf{Q}$-linear combination of characters induced by trivial representations of cyclic subgroups, one shows that $\chi_c(Y,\alpha) \in \mathrm{K}_0^{mot}(\mathrm{Var}_k) \otimes \mathbf{Q}$. For $X := Y/G$ and C any cyclic subgroup of G, we define $\chi_c([\varphi_{Y,X,C}]) := \chi_c(Y,\theta)$, where θ sends $g \in G$ to 1 if the subgroup generated by g is conjugate to C, and else to 0. Note that θ equals $|C| \,/\, |N_G(C)|$ times the function on G induced by the characteristic function on C of the set of generators of C. This implies our requirement (ii), because of Proposition 3.1.2.(2) of [10]. The map $(Y,\alpha) \mapsto \chi_c(Y,\alpha)$ satisfies the nice compatibility relations stated in Proposition 3.1.2 of loc. cit. This compatibility (together with the above mentioned quantifier elimination) is used, exactly as in loc. cit., to prove that the above definition of $\chi_c([\varphi_{Y,X,C}])$ extends by additivity to a well-defined map $\chi_c : \mathrm{K}_0(\mathrm{PFF}_k) \longrightarrow \mathrm{K}_0^{mot}(\mathrm{Var}_k) \otimes \mathbf{Q}$. In loc. cit., Chow motives with coefficients in the algebraic closure of $\mathbf{Q}$ are used, but we can work as well with coefficients in $\mathbf{Q}$, since here we only have to consider representations of G over $\mathbf{Q}$.

3. Arc spaces and the motivic Poincaré series

Let X be an algebraic variety defined over a field k of characteristic zero. For any natural number n, the n-th jet space $\mathcal{L}_n(X)$ of X is the unique algebraic variety over k whose K-rational points correspond in a bijective and functorial way to the rational points on X over $K[t]/t^{n+1}$, for any field K containing k. The arc space $\mathcal{L}(X)$ of X is the reduced k-scheme obtained by taking the projective limit of the varieties $\mathcal{L}_n(X)$ in the category of k-schemes.

We will now give the definition of the motivic Poincaré series $P(T)$ of X. This series is called the arithmetic Poincaré series in [10], and is very different from the geometric Poincaré series studied in [8]. For notational convenience we only give the definition here when X is a subvariety of some affine space $\mathbf{A}_k^m$. For the general case we refer to section 5 below or to our paper [10]. By Greenberg's Theorem [16], for each n there exists a ring formula φ_n over k such that, for all fields K containing k, the K-rational points of $\mathcal{L}_n(X)$, that can be lifted to a K-rational point of $\mathcal{L}(X)$, correspond to the tuples satisfying the interpretation of φ_n in K. (The correspondence is induced by mapping a polynomial over K to the tuple consisting of its coefficients.) Clearly, when two formulas satisfy this requirement, then they are equivalent when interpreted in any field containing k, and hence define the same class in $\mathrm{K}_0(\mathrm{PFF}_k)$. Now we are ready to give the definition of $P(T)$:

$$P(T) := \sum_n \chi_c([\varphi_n])T^n.$$

Theorem 3.1. *The motivic Poincaré series $P(T)$ is a rational power series over the ring $\mathrm{K}_0^{mot}(\mathrm{Var}_k)[\mathbf{L}^{-1}] \otimes \mathbf{Q}$, with denominator a product of factors of the form $1 - \mathbf{L}^a T^b$, with $a, b \in \mathbf{Z}$, $b > 0$. Moreover if $k = \mathbf{Q}$, the Serre Poincaré*

series $P_p(T)$, for $p \gg 0$, is obtained from $P(T)$ by applying the operator N_p to each coefficient of the numerator and denominator of $P(T)$.

In particular we see that the degrees of the numerator and the denominator of $P_p(T)$ remain bounded for p going to infinity. This fact was first proved by Macintyre [23] and Pas [26].

4. Quantifier elimination for valuation rings

Let R be a ring and assume it is an integral domain. We will define the notion of a DVR-formula over R. Such a formula can be interpreted in any discrete valuation ring $A \supset R$ with a distinguished uniformizer π. It can contain variables that run over the discrete valuation ring, variables that run over the value group $\mathbf{Z}$, and variables that run over the residue field. A DVR-formula over R is build from quantifiers with respect to variables that run over the discrete valuation ring, or over the value group, or over the residue field, Boolean combinations, and expressions of the following form: $g_1(x) = 0$, $\mathrm{ord}(g_1(x)) \leq \mathrm{ord}(g_2(x)) + L(a)$, $\mathrm{ord}(g_1(x)) \equiv L(a) \bmod d$, where $g_1(x)$ and $g_2(x)$ are polynomials over R in several variables x running over the discrete valuation ring, where $L(a)$ is a polynomial of degree ≤ 1 over $\mathbf{Z}$ in several variables a running over the value group, and d is any positive integer (not a variable). Moreover we also allow expressions of the form $\varphi(\overline{\mathrm{ac}}(h_1(x)), ..., \overline{\mathrm{ac}}(h_r(x)))$, where φ is a ring formula over R, to be interpreted in the residue field, $h_1(x), ..., h_r(x)$ are polynomials over R in several variables x running over the discrete valuation ring, and $\overline{\mathrm{ac}}(v)$, for any element v of the discrete valuation ring, is the residue of the angular component $\mathrm{ac}(v) := v\pi^{-\mathrm{ord}\, v}$. For the discrete valuation rings $\mathbf{Z}_p$ and $K[[t]]$, we take as distinguished uniformizer π the elements p and t.

Theorem 4.1 (Quantifier Elimination of Pas [26]). *Suppose that R has characteristic zero. For any DVR-formula θ over R there exists a DVR-formula ψ over R, which contains no quantifiers running over the valuation ring and no quantifiers running over the value group, such that*
(1) $\theta \longleftrightarrow \psi$ holds in $K[[t]]$, for all fields K containing R.
(2) $\theta \longleftrightarrow \psi$ holds in $\mathbf{Z}_p$, for all primes $p \gg 0$, when $R = \mathbf{Z}$.

The Theorem of Pas is one of several quantifier elimination results for Henselian valuation rings, and goes back to the work of Ax-Kochen-Ersov and Cohen on the model theory of valued fields, which was further developed by Macintyre, Delon [6], and others, see e.g. Macintyre's survey [22].

Combining the Theorem of Pas with the work of Ax mentioned in section 2, one obtains

Theorem 4.2 (Ax-Kochen-Ersov Principle, version of Pas). *Let σ be a DVR-formula over $\mathbf{Z}$ with no free variables. Then the following are equivalent:*
(i) The interpretation of σ in $\mathbf{Z}_p$ is true for all primes $p \gg 0$.
(ii) The interpretation of σ in $K[[t]]$ is true for all pseudo-finite fields K of characteristic zero.

5. Definable subassignements and truncations

Let $h : \mathcal{C} \to$ Sets be a functor from a category $\mathcal{C}$ to the category of sets. We shall call the data for each object C of $\mathcal{C}$ of a subset $h'(C)$ of $h(C)$ a *subassignement* of h. The point in this definition is that h' is not assumed to be a subfunctor of h. For h' and h'' two subassignements of h, we shall denote by $h' \cap h''$ and $h' \cup h''$, the subassignements $C \mapsto h'(C) \cap h''(C)$ and $C \mapsto h'(C) \cup h''(C)$, respectively.

Let k be a field of characteristic zero. We denote by Field_k the category of fields which contain k. For X a variety over k, we consider the functor $h_X : K \mapsto X(K)$ from Field_k to the category of sets. Here $X(K)$ denotes the set of K-rational points on X. When X is a subvariety of some affine space, then a subassignement h of h_X is called *definable* if there exists a ring formula φ over k such that, for any field K containing k, the set of tuples that satisfy the interpretation of φ in K, equals $h(K)$. Moreover we define the *class* $[h]$ *of* h in $\mathrm{K}_0(\mathrm{PFF}_k)$ as $[\varphi]$. More generally, for any algebraic variety X over k, a subassignement h of h_X is called *definable* if there exists a finite cover $(X_i)_{i\in I}$ of X by affine open subvarieties and definable subassignements h_i of h_{X_i}, for $i \in I$, such that $h = \cup_{i\in I} h_i$. The *class* $[h]$ *of* h in $\mathrm{K}_0(\mathrm{PFF}_k)$ is defined by linearity, reducing to the affine case.

For any algebraic variety X over k we denote by $h_{\mathcal{L}(X)}$ the functor $K \mapsto X(K[[t]])$ from Field_k to the category of sets. Here $X(K[[t]])$ denotes the set of $K[[t]]$-rational points on X. When X is a subvariety of some affine space, then a subassignement h of $h_{\mathcal{L}(X)}$ is called *definable* if there exists a DVR-formula φ over k such that, for any field K containing k, the set of tuples that satisfy the interpretation of φ in $K[[t]]$, equals $h(K)$. More generally, for any algebraic variety X over k, a subassignement h of $h_{\mathcal{L}(X)}$ is called *definable* if there exists a finite cover $(X_i)_{i\in I}$ of X by affine open subvarieties and definable subassignements h_i of $h_{\mathcal{L}(X_i)}$, for $i \in I$, such that $h = \cup_{i\in I} h_i$. A family of definable subassignements h_n, $n \in \mathbf{Z}$, of $h_{\mathcal{L}(X)}$ is called a *definable family of definable subassignements* if on each affine open of a suitable finite affine covering of X, the family h_n is given by a DVR-formula containing n as a free variable running over the value group.

Let X be a variety over k. Let h be a definable subassignement of $h_{\mathcal{L}(X)}$, and n a natural number. The *truncation of* h *at level* n, *denoted by* $\pi_n(h)$, is the subassignement of $h_{\mathcal{L}_n(X)}$ that associates to any field K containing k the image of $h(K)$ under the natural projection map from $X(K[[t]])$ to $\mathcal{L}_n(X)(K)$. Using the Quantifier Elimination Theorem of Pas, we proved that $\pi_n(h)$ is a definable subassignement of $h_{\mathcal{L}_n(X)}$, so that we can consider its class $[\pi_n(h)]$ in $\mathrm{K}_0(\mathrm{PFF}_k)$. Using the notion of truncations, we can now give an alternative (but equivalent) definition of the motivic Poincaré series $P(T)$, which works for any algebraic variety X over k, namely $P(T) := \sum_n \chi_c([\pi_n(h_{\mathcal{L}(X)})])T^n$.

A definable subassignement h of $h_{\mathcal{L}(X)}$ is called *weakly stable at level* n if for any field K containing k the set $h(K)$ is a union of fibers of the natural projection map from $X(K[[t]])$ to $\mathcal{L}_n(X)(K)$. If X is nonsingular, with all its irreducible components of dimension d, and h is a definable subassignement of $h_{\mathcal{L}(X)}$, which is weakly stable at level n, then it is easy to verify that

$$[\pi_n(h)]\mathbf{L}^{-nd} = [\pi_m(h)]\mathbf{L}^{-md}$$

for all $m \geq n$. Indeed this follows from the fact that the natural map from $\mathcal{L}_m(X)$ to $\mathcal{L}_n(X)$ is a locally trivial fibration for the Zariski topology with fiber $\mathbf{A}_k^{(m-n)d}$, when X is nonsingular.

6. Arithmetic motivic integration

Here we will outline an extension of the theory of motivic integration, called arithmetic motivic integration. If the base field k is algebraically closed, then it coincides with the usual motivic integration.

We denote by $\widehat{\mathrm{K}}_0^{mot}(\mathrm{Var}_k)[\mathbf{L^{-1}}]$ the completion of $\mathrm{K}_0^{mot}(\mathrm{Var}_k)[\mathbf{L^{-1}}]$ with respect to the filtration of $\mathrm{K}_0^{mot}(\mathrm{Var}_k)[\mathbf{L^{-1}}]$ whose m-th member is the subgroup generated by the elements $[V]\mathbf{L}^{-i}$ with $i - \dim V \geq m$. Thus a sequence $[V_i]\mathbf{L}^{-i}$ converges to zero in $\widehat{\mathrm{K}}_0^{mot}(\mathrm{Var}_k)[\mathbf{L^{-1}}]$, for $i \mapsto +\infty$, if $i - \dim V_i \mapsto +\infty$.

Definition-Theorem 6.1. *Let X be an algebraic variety of dimension d over a field k of characteristic zero, and let h be a definable subassignement of $h_{\mathcal{L}(X)}$. Then the limit*

$$\nu(h) := \lim_{n\to\infty} \chi_c([\pi_n(h)])\mathbf{L}^{-(n+1)d}$$

exists in $\widehat{\mathrm{K}}_0^{mot}(\mathrm{Var}_k)[\mathbf{L}^{-1}] \otimes \mathbf{Q}$ and is called the arithmetic motivic volume of h.

We refer to [10, §6] for the proof of the above theorem. If X is nonsingular and h is weakly stable at some level, then the theorem follows directly from what we said at the end of the previous section. When X is nonsingular affine, but h general, the theorem is proved by approximating h by definable subassignements h_i of $h_{\mathcal{L}(X)}$, $i \in \mathbf{N}$, which are weakly stable at level $n(i)$. For h_i we take the subassignement obtained from h by adding, in the DVR-formula φ defining h, the condition $\mathrm{ord}g(x) \leq i$, for each polynomial $g(x)$ over the valuation ring, that appears in φ. (Here we assume that φ contains no quantifiers over the valuation ring.) It remains to show then that $\chi_c([\pi_n(\mathrm{ord}g(x) > i)])\mathbf{L}^{-(n+1)d}$ goes to zero when both i and $n \gg i$ go to infinity, but this is easy.

Theorem 6.2. *Let X be an algebraic variety of dimension d over a field k of characteristic zero, and let h, h_1 and h_2 be definable subassignements of $h_{\mathcal{L}(X)}$.*
(1) If $h_1(K) = h_2(K)$ for any pseudo-finite field $K \supset k$, then $\nu(h_1) = \nu(h_2)$.
(2) $\nu(h_1 \cup h_2) = \nu(h_1) + \nu(h_2) - \nu(h_1 \cap h_2)$
(3) If S is a subvariety of X of dimension $< d$, and if $h \subset h_{\mathcal{L}(S)}$, then $\nu(h) = 0$.
(4) Let h_n, $n \in \mathbf{N}$, be a definable family of definable subassignements of $h_{\mathcal{L}(X)}$. If $h_n \cap h_m = \emptyset$, for all $n \neq m$, then $\sum_n \nu(h_n)$ is convergent and equals $\nu(\bigcup_n h_n)$.
(5) Change of variables formula. *Let $p : Y \to X$ be a proper birational morphism of nonsingular irreducible varieties over k. Assume for any field K containing k that the jacobian determinant of p at any point of $p^{-1}(h(K))$ in $Y(K[[t]])$ has t-order equal to e. Then $\nu_X(h) = \mathbf{L}^{-e}\nu_Y(p^{-1}(h))$. Here ν_X, ν_Y denote the arithmetic motivic volumes relative to X, Y, and $p^{-1}(h)$ is the subassignement of $h_{\mathcal{L}(Y)}$ given by $K \mapsto p^{-1}(h(K)) \cap Y(K[[t]])$.*

Assertion (1) is a direct consequence of the definitions. Assertions (2) and (4) are proved by approximating the subassignements by weakly stable ones. Moreover

for (4) we also need the fact that $h_n = \emptyset$ for all but a finite number of n's, when all the h_n, and their union, are weakly stable (at some level depending on n). Assertion (5) follows from the fact that for $n \gg e$ the map $\mathcal{L}_n(Y) \to \mathcal{L}_n(X)$ induced by p is a piecewise trivial fibration with fiber $\mathbf{A}_k^e$ over the image in $\mathcal{L}_n(X)$ of the points of $\mathcal{L}(Y)$ where the jacobian determinant of p has t-order e. See [10] for the details.

7. About the proof of Theorem 3.1

We give a brief sketch of the proof of Theorem 3.1, in the special case that X is a hypersurface in $\mathbf{A}_k^d$ with equation $f(x) = 0$. Actually, here we will only explain why the image $\widehat{P}(T)$ of $P(T)$ in the ring of power series over $\widehat{\mathrm{K}}_0^{mot}(\mathrm{Var}_k)[\mathbf{L}^{-1}] \otimes \mathbf{Q}$ is rational. The rationality of $P(T)$ requires additional work. Let $\varphi(x, n)$ be the DVR-formula $(\exists y)(f(y) = 0$ and $\mathrm{ord}(x - y) \geq n)$, with d free variables x running over the discrete valuation ring, and one free variable n running over the value group. That formula determines a definable family of definable subassignements $h_{\varphi(-,n)}$ of $h_{\mathcal{L}(\mathbf{A}_k^d)}$. Since $h_{\varphi(-,n)}$ is weakly stable at level n, unwinding our definitions yields that the arithmetic motivic volume on $h_{\mathcal{L}(\mathbf{A}_k^d)}$ of $h_{\varphi(-,n)}$ equals $\mathbf{L}^{-(n+1)d}$ times the n-th coefficient of $\widehat{P}(T)$. To prove that $\widehat{P}(T)$ is a rational power series we have to analyze how the arithmetic motivic volume of $h_{\varphi(-,n)}$ depends on n. To study this, we use Theorem 4.1 (quantifier elimination of Pas) to replace the formula $\varphi(x, n)$ by a DVR-formula $\psi(x, n)$ with no quantifiers running over the valuation ring and no quantifiers over the value group. We take an embedded resolution of singularities $\pi : Y \longrightarrow \mathbf{A}_k^d$ of the union of the loci of the polynomials over the valuation ring, that appear in $\psi(x, n)$. Thus the pull-backs to Y of these polynomials, and the jacobian determinant of π, are locally a monomial times a unit. Thus the pull-back of the formula $\psi(x, n)$ is easy to study, at least if one is not scared of complicated formula in residue field variables. The key idea is to calculate the arithmetic motivic volume of $h_{\psi(-,n)}$, by expressing it as a sum of arithmetic motivic volumes on $h_{\mathcal{L}(Y)}$, using the change of variables formula in Theorem 6.2. These volumes can be computed explicitly, and this yields the rationality of $\widehat{P}(T)$.

To prove that $\widehat{P}(T)$ specializes to the Serre Poincaré series $P_p(T)$ for $p \gg 0$, we repeat the above argument working with $\mathbf{Z}_p^d$ instead of $\mathcal{L}(\mathbf{A}_k^d)$. The p-adic volume of the subset of $\mathbf{Z}_p^d$ defined by the formula $\varphi(x, n)$ equals $p^{-(n+1)d}$ times the n-th coefficient of $P_p(T)$. Because of Theorem 4.1.(2), we can again replace $\varphi(x, n)$ by the formula $\psi(x, n)$ that we obtained already above. That p-adic volume can be calculated explicitly by pulling it back to the p-adic manifold $Y(\mathbf{Z}_p)$, and one verifies a posteriory that it is obtained by applying the operator N_p to the arithmetic motivic volume that we calculated above. This verification uses the last assertion in Theorem 2.1.

8. The general setting

We denote by $\overline{\mathcal{M}}$ the image of $\mathrm{K}_0^{mot}(\mathrm{Var}_k)[\mathbf{L}^{-1}]$ in $\widehat{\mathrm{K}}_0^{mot}(\mathrm{Var}_k)[\mathbf{L}^{-1}]$, and by $\overline{\mathcal{M}}_{loc}$ the localization of $\overline{\mathcal{M}} \otimes \mathbf{Q}$ obtained by inverting the elements $\mathbf{L}^i - 1$, for all

$i \geq 1$. One verifies that the operator N_p can be applied to any element of $\overline{\mathcal{M}}_{loc}$, for $p \gg 0$, yielding a rational number. The same holds for the Hodge-Deligne polynomial which now belongs to $\mathbf{Q}(u,v)$. By the method of section 7, we proved in [10] the following

Theorem 8.1. *Let X be an algebraic variety over a field k of characteristic zero, let h be a definable subassignement of $h_{\mathcal{L}(X)}$, and h_n a definable family of definable subassignements of $h_{\mathcal{L}(X)}$.*

(1) The motivic volume $\nu(h)$ is contained in $\overline{\mathcal{M}}_{loc}$.

(2) The power series $\sum_n \nu(h_n)T^n \in \overline{\mathcal{M}}_{loc}[[T]]$ is rational, with denominator a product of factors of the form $1 - \mathbf{L}^{-a}T^b$, with $a, b \in \mathbf{N}$, $b \neq 0$.

Let X be a reduced separable scheme of finite type over $\mathbf{Z}$, and let $A = (A_p)_{p \gg 0}$ be a definable family of subsets of $X(\mathbf{Z}_p)$, meaning that on each affine open, of a suitable finite affine covering of X, A_p can be described by a DVR-formula over $\mathbf{Z}$. (Here p runs over all large enough primes.) To A we associate in a canonical way, its motivic volume $\nu(h_A) \in \overline{\mathcal{M}}_{loc}$, in the following way: Let h_A be a definable subassignement of $h_{\mathcal{L}(X \otimes \mathbf{Q})}$, given by DVR-formulas that define A. Because these formulas are not canonical, the subassignement h_A is not canonical. But by the Ax-Kochen-Ersov Principle (see 4.2), the set $h_A(K)$ is canonical for each pseudo-finite field K containing $\mathbf{Q}$. Hence $\nu(h_A) \in \overline{\mathcal{M}}_{loc}$ is canonical, by Theorem 6.2.(1). By the method of section 7, we proved in [10] the following comparison result:

Theorem 8.2. *With the above notation, for all large enough primes p, $N_p(\nu(h_A))$ equals the measure of A_p with respect to the canonical measure on $X(\mathbf{Z}_p)$.*

When $X \otimes \mathbf{Q}$ is nonsingular and of dimension d, the canonical measure on $X(\mathbf{Z}_p)$ is defined by requiring that each fiber of the map $X(\mathbf{Z}_p) \to X(\mathbf{Z}_p/p^m)$ has measure p^{-md} whenever $m \gg 0$. For the definition of the canonical measure in the general case, we refer to [25].

The above theorem easily generalizes to integrals instead of measures, but this yields little more because quite general p-adic integrals (such as the orbital integrals appearing in the Langlands program) can be written as measures of the definable sets we consider. For example the p-adic integral $\int |f(x)|dx$ on $\mathbf{Z}_p^d$ equals the p-adic measure of $\{(x,t) \in \mathbf{Z}_p^{d+1} : \mathrm{ord}_p(f(x)) \leq \mathrm{ord}_p(t)\}$.

References

[1] J. Ax, The elementary theory of finite fields, *Ann. of Math*, 88 (1968), 239–271.

[2] S. del Baño Rollin, V. Navarro Aznar, On the motive of a quotient variety, *Collect. Math.*, 49 (1998), 203–226.

[3] V. Batyrev, Stringy Hodge numbers of varieties with Gorenstein canonical singularities, *Integrable systems and algebraic geometry (Kobe/Kyoto, 1997)*, 1–32, World Sci. Publishing, River Edge, NJ, 1998.

[4] V, Batyrev, Non-Archimedean integrals and stringy Euler numbers of log-terminal pairs, *J. Eur. Math. Soc. (JEMS)*, 1 (1999), 5–33.

[5] Z. Chatzidakis, L. van den Dries, A. Macintyre, Definable sets over finite fields, *J. Reine Angew. Math.*, 427 (1992), 107-135.

[6] F. Delon, *Quelques propriétés des corps valués*, Thèse d' État, Université Paris VII (1981).
[7] J. Denef, The rationality of the Poincaré series associated to the p-adic points on a variety, *Invent. Math.*, 77 (1984), 1–23.
[8] J. Denef, F. Loeser, Germs of arcs on singular algebraic varieties and motivic integration, *Invent. Math.*, 135 (1999), 201–232.
[9] J. Denef, F. Loeser, Motivic exponential integrals and a motivic Thom-Sebastiani theorem, *Duke Math. J.*, 99 (1999), no. 2, 285–309.
[10] J. Denef, F. Loeser, Definable sets, motives and p-adic integrals, *J. Amer. Math. Soc.*, 14 (2001), 429–469.
[11] J. Denef, F. Loeser, Geometry on arc spaces of algebraic varieties, *Proceedings of the Third European Congress of Mathematics*, Volume 1, Progress in Mathematics 201, Birkhauser 2001, ISBN 3-7643-6417-3.
[12] J. Denef, F. Loeser, Motivic integration, quotient singularities and the McKay correspondence, *Compositio Math.*, 131 (2002), 267–290.
[13] M. Fried, M. Jarden, *Field arithmetic*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), Springer-Verlag, Berlin, 1986. ISBN: 3-540-16640-8.
[14] M. Fried, G. Sacerdote, Solving diophantine problems over all residue class fields of a number field and all finite fields, *Ann. Math.*, 100 (1976), 203–233.
[15] H. Gillet, C. Soulé, Descent, motives and K-theory, *J. Reine Angew. Math.*, 478 (1996), 127–176.
[16] M. Greenberg, Rational points in Henselian discrete valuation rings, *Inst. Hautes Études Sci. Publ. Math.*, 31 (1966), 59–64.
[17] F. Guillén, V. Navarro Aznar, Un critère d'extension d'un foncteur défini sur les schémas lisses, preprint (1995), revised (1996).
[18] T.C. Hales, Can p-adic integrals be computed? (to appear).
[19] C. Kiefe, Sets definable over finite fields: their zeta-functions, *Trans. Amer. Math. Soc.*, 223 (1976), 45–59.
[20] M. Kontsevich, Lecture at Orsay (December 7, 1995).
[21] E. Looijenga, Motivic measures, *Séminaire Bourbaki*, Vol. 1999/2000, Astérisque 276 (2002), 267–297.
[22] A. Macintyre, Twenty years of p-adic model theory, *Logic colloquium '84*, 121–153, Stud. Logic Found. Math., 120, North-Holland, Amsterdam, 1986.
[23] A. Macintyre, Rationality of p-adic Poincaré series: Uniformity in p, *Ann. Pure Appl. Logic*, 49 (1990), 31–74.
[24] M. Mustaţă, Jet schemes of locally complete intersection canonical singularities, *Invent. Math.*, 145 (2001), 397–424.
[25] J. Oesterlé, Réduction modulo p^n des sous-ensembles analytiques fermés de $\mathbf{Z}_p^N$, *Invent. Math.*, 66 (1982), 325–341.
[26] J. Pas, Uniform p-adic cell decomposition and local zeta functions, *J. Reine Angew. Math.*, 399 (1989) 137–172.
[27] A.J. Scholl, Classical motives, *Motives*, Seattle, WA, 1991, 163–187, Proc. Sympos. Pure Math., 55, Part 1, Amer. Math. Soc., Providence, RI, 1994.

ICM 2002 · Vol. II · 25–33

Automorphism Groups of Saturated Structures; A Review

D. Lascar*

Abstract

We will review the main results concerning the automorphism groups of saturated structures which were obtained during the two last decades. The main themes are: the small index property in the countable and uncountable cases; the possibility of recovering a structure or a significant part of it from its automorphism group; the subgroup of strong automorphisms.

2000 Mathematics Subject Classification: 03C50, 20B27.
Keywords and Phrases: Automorphism groups, Small index property, Strong automorphisms.

1. Introduction

Saturated models play an important role in model theory. In fact, when studying the model theory of a complete theory T, one may work in a large saturated model of T with its definable sets, and forget everything else about T. This large saturated structure is sometimes called the "universal domain", sometimes the "monster model".

A significant work has been done the last twenty years on the automorphism groups of saturated models. It is this work that I want to review here. There is a central question that I will use as a "main theme" to organize the paper: what information about M and its theory are contained in its group of automorphisms? In the best case, M itself is "encoded" in some way in this group; recovering M from it is known as "the reconstruction problem". A possible answer to this problem is a theorem of the form: If M_1 and M_2 are structures in a given class with isomorphic automorphism groups, then M_1 and M_2 are isomorphic.

Throughout this paper, T is supposed to be a countable complete theory. The countability of T is by no means an essential hypothesis. Its purpose is only to make the exposition smoother, and most of the results generalize without difficulty

*CNRS, Université Denis Diderot Paris 7, 2 Place Jussieu, UFR de mathématiques, case 7012, 75521 Paris Cedex 05, France. E-mail: lascar@logique.jussieu.fr

to uncountable theories. We will denote by $Aut(M)$ the group of automorphisms of the structure M, and if A is a subset of M, $Aut_A(M)$ will be the pointwise stabilizer of A:

$$Aut_A(M) = \{f \in Aut(M) \; ; \; \forall a \in A \, f(a) = a\} \, .$$

When we say "definable", we mean "definable without parameters".

2. The countable case

As a preliminary remark, let us say that the automorphism group of a saturated model is always very rich: if M has cardinality λ, then its automorphism group has cardinality 2^λ.

I do not know who was the first to introduce the small index property. As we will see, it is crucial in the subject.

Definition 1 *Let M be a countable structure. We say that M (or $Aut(M)$) has the small index property if for any subgroup H of $Aut(M)$ of index less than $2^{\aleph_0}$, there exists a finite set $A \subset M$ such that $Aut_A(M) \subseteq H$.*

Remark that the converse is true: any subgroup containing a subgroup of the form $Aut_A(M)$ where A is finite, has a countable index in $Aut(M)$. Moreover, the subgroups containing a subgroup of the form $Aut_A(M)$ are precisely the open neighborhoods of the identity for the pointwise convergence topology. In other words, the small index property allows us to recover the topological structure of $Aut(M)$ from its pure group structure.

The small index property has been proved for a number of countable saturated structures:

1. The infinite set without additional structure [23], [5].
2. The linear densely ordered sets [25].
3. The vector spaces over a finite field [6].
4. The random graph [10].
5. Various other classes of graphs [9].
6. Generic relational structures [8].
7. ω-categorical ω-stable structures [10].

The small index property has also been proved for some countable structures which are not saturated: for the free group with ω-generators ([2]), for arithmetically saturated models of arithmetic ([17]).

There are examples of countable saturated structures which fail to have the small index property. The simplest may be an algebraically closed field of characteristic 0 of infinite countable transcendence degree: Let $\bar{\mathbb{Q}}$ be the algebraic closure of the field of rational numbers. There is an obvious homomorphism φ from $Aut(M)$ onto $Aut(\bar{\mathbb{Q}})$ (the restriction map). Now, it is well known that there is a subgroup H of $Aut(\bar{\mathbb{Q}})$ of countable index (in fact of finite index) which is not closed for the Krull topology, which is nothing else that the pointwise convergence topology. Then $\varphi^{-1}(H)$ is not open, but of finite index in $Aut(M)$.

As we will see later, the small index property is particularly relevant for ω-categorical structures. Evans and Hewitt have produced an example of such a structure without the small index property ([7]).

With the pointwise convergence topology, $Aut(M)$ is a topological polish group. So, we may use the powerful tools of descriptive set theory. In many cases (for example for structures 1-6 above), it can be shown that there is a (necessarily unique) conjugacy class which is generic, that is, is the countable intersection of dense open subsets. The elements of this class are called generic automorphisms, and they play an important role in the proof of the small index property.

Another possible nice property of these automorphism groups which is sometimes obtained as a bonus of the proof of the small index property, is the fact that its cofinality is not countable, that is, $Aut(M)$ is not the union of a countable chain of proper subgroups. This is proved in particular for the full permutation group of a countable set ([21]), for the random graph and for ω-categorical ω-stable structures ([10]).

I would like to mention here the work of Rubin ([24]). He has shown how to reconstruct a certain number of structures from their automorphism group using a somewhat different method. His methods apply essentially to "combinatorial structures" such as the random graph, the universal homogeneous poset, the generic tournament (a structure for which the small index property is not known), etc.

3. Subgroups and imaginary elements

Recall that an imaginary element of M is a class of a tuple of M^n modulo a definable equivalence relation on M^n. For instance, if G is a group and H a definable subgroup of G^n, then any coset of H in G^n is an imaginary element. When we add all these imaginary elements to a saturated structure M, we obtain the structure M^{eq}, and we can consider M^{eq} as a saturated structure (in a larger language).

It is clear that M and M^{eq} have canonically the same automorphism group: every automorphism of M extends uniquely to an automorphism of M^{eq}. This shows a limitation to the reconstruction problem: If M and N are two structures which are such that "M^{eq} and N^{eq} are isomorphic", then $Aut(M)$ and $Aut(N)$ are isomorphic via a bicontinuous isomorphism. The condition "M^{eq} and N^{eq} are isomorphic" may seem weird, but in fact, it is natural. Roughly speaking, it means that M can be interpreted in N, and conversely (a little more in fact, see [1] for more details). In this case, we say that M and N are bi-interpretable.

Consider now the case of an ω-categorical structure M. It is not difficult to see that any open subgroup of $Aut(M)$ is the stabilizer $Aut_\alpha(M)$ of an imaginary element α. Moreover, $Aut(M)$ acts by conjugation on the set of its open subgroups, and this action is (almost) isomorphic to the action of $Aut(M)$ on M^{eq} (almost because two different imaginary elements α and β may have the same stabilizer). So, from the topological group $Aut(M)$ we can (almost) reconstruct its action on M^{eq}. We can do better:

Theorem 2 [1] *Assume that M and N are countable ω-categorical structures. Then the following two conditions are equivalent:*

1. *there is a bicontinuous isomorphism from $Aut(M)$ onto $Aut(N)$;*
2. *M and N are bi-interpretable.*

In fact, these conditions are also equivalent to: there exists a continuous isomorphism from $Aut(M)$ onto $Aut(N)$ (see [15]). Thus, if one of the structure M or N has the small index property and $Aut(M)$ is isomorphic to $Aut(N)$ (as pure groups), then M and N are bi-interpretable.

Now, if M is not necessarily ω-categorical (but still saturated), the situation is a bit more complicated. We need to introduce new elements.

Definition 3 *1. An ultra-imaginary element of M is a class modulo E, where E is an equivalence relation on M^n ($n \leq \omega$) which is invariant under the action of $Aut(M)$. An ultra-imaginary element is finitary if $n < \omega$.*
2. A hyperimaginary element of M is a class modulo E, where E is an equivalence relation on M^n ($n \leq \omega$) which is defined by a (possibly infinite) conjunction of first order formulas.

An imaginary element is hyperimaginary, and a hyperimaginary element is ultra-imaginary. A hyperimaginary element is a class modulo an equivalence relation E defined by a formula of the form

$$\bigwedge_{i \in I} \varphi_i$$

where the φ_i are first-order formulas (without parameters) and whose free variables are among the x_k for $k < n$. An ultra-imaginary element is a class modulo an equivalence relation E defined by a formula of the form

$$\bigvee_{j \in J} \bigwedge_{i \in I} \varphi_{i,j}$$

where the $\varphi_{i,j}$ are first order-formulas (without parameters) and whose free variables are among the x_k for $k < n$.

If M is a countable saturated structure, the stabilizer of a finitary ultra-imaginary element is clearly an open subgroup, and it is not difficult to see that if H an open subgroup of $Aut(M)$, then there exists a finitary ultra-imaginary element α such that H is the stabilizer of α. In the ω-categorical case, any finitary ultra-imaginary is in fact imaginary, and this explain why this case is so simple.

In some cases, for example for ω-stable theories (see [18]), it is possible to characterize, among all open subgroups, those which are of the form $Aut_\alpha(M)$ with α imaginary. Something similar has been done for countable arithmetically saturated models of arithmetic in [11], and in [13], it is proved that if two such models have isomorphic automorphism groups, then they are isomorphic.

4. Strong automorphisms

It is now time to introduce the group of strong automorphisms.

Definition 4 [14] *The group of strong automorphisms of M is the group generated by the set*

$$\bigcup \{Aut_N(M) \ ; \ N \prec M\}$$

and is denoted $Autf(M)$.

It is easy to see that $Autf(M)$ is a normal subgroup of $Aut(M)$. Its index is at most $2^{\aleph_0}$. Moreover, the quotient group $Aut(M)/Autf(M)$ depend only on T : if M and M' are two saturated models, $M \prec M'$, then there is a natural isomorphism from $Aut(M)/Autf(M)$ onto $Aut(M')/Autf(M')$. $Aut(M)/Autf(M)$ will be denoted $Gal(T)$ (of course, Gal stands for Galois). For example, if T is the theory of algebraically closed fields of characteristic 0, $Autf(M) = Aut_{\bar{\mathbb{Q}}}(M)$ and $Gal(T)$ is (isomorphic to) the group of automorphisms of $\bar{\mathbb{Q}}$.

In fact this interpretation is general. Assume first that M is of cardinality bigger than $2^{\aleph_0}$. Let α be an ultra-imaginary element of M. It can be shown that the following conditions are equivalent:

1. $card\{f(\alpha) \ ; \ f \in Aut(M)\} < card(M)$;
2. $card\{f(\alpha) \ ; \ f \in Aut(M)\} \leq 2^{\aleph_0}$;
   An equivalence relation is bounded if it has at most $2^{\aleph_0}$ classes (equivalently less than $card(M)$ classes). The above conditions are also equivalent to:
3. α (as a set) is the class modulo an invariant bounded equivalence relation.

If these conditions are satisfied, we say that α is bounded. It should be remarked that an imaginary element is bounded if and only if it is algebraic, if and only if its orbit is finite.

We will denote by $Bdd(M)$ the set of bounded ultra-imaginary elements of M. This set does not really depend on M (as soon as its cardinality is big enough) but only on its theory: any invariant bounded equivalence relation has a representative in any uncountable saturated model. We will allow ourself to write $Bdd(T)$ when convenient. Moreover $Autf(M)$ is exactly the pointwise stabilizer of $Bdd(M)$ so that $Gal(T)$ can be identified with the group of elementary permutations of $Bdd(T)$.

With some care, we can generalize this interpretation to models of small cardinality: for example, assume M countable, and let M' be a large saturated extension of M. Then any automorphism f of M extends to an automorphism of M', and if f_1 and f_2 are two such extensions, then their action on $Bdd(M')$ are equal; $Autf(M)$ is exactly the set of automorphisms whose extensions to M' act trivially on $Bdd(M')$.

In any case, $Aut(M)$ leaves fixed the set of bounded imaginary elements and the set of bounded hyperimaginary elements. In some cases (for example for algebraically closed fields), $Gal(T)$ acts faithfully on the set of bounded imaginary elements. It is the case if T is stable ([14]). It is not known if it is always true for simple theories, but it is true for the so-called low simple theories ([3]) and in particular for supersimple theories. For simple theories, $Gal(T)$ acts faithfully on the set of bounded hyperimaginary elements ([12]). In [4] there is an example of a theory where the action of $Gal(T)$ on the set of hyperimaginary elements is not faithful.

There is a natural topology on $Gal(T)$ (see [19] for details). It can be defined in two different ways.

My favorite one is via the ultraproduct construction. Let $(\gamma_i\ ;\ i \in I)$ be a family of elements of $Gal(T)$ and $\mathcal{U}$ an ultrafilter on I. Choose a saturated model M and, for each $i \in I$ an automorphism $f_i \in Aut(M)$ lifting γ_i. Consider the ultrapower $M' = \prod_{i\in\mathcal{U}} M$. We can define the automorphism $\prod_{i\in\mathcal{U}} f_i$ on M'. This automorphism acts on $Bdd(M') = Bdd(T)$, so defines an element of $Gal(T)$, say β. This element β should be considered as a limit of the family $(\gamma_i\ ;\ i \in I)$ along $\mathcal{U}$. A subset X of $Gal(T)$ is closed for the topology we are defining if it is closed for this limit operation. You should be aware that the element β may depend on the choices of the f_i's, because the topology we are defining is not Hausdorff in general.

The other way to define a topological structure on $Gal(T)$ is to define a topology on $Bdd(T)$. If, as it is the case when T is stable, $Gal(T)$ can be identified with a group of permutation on the set of imaginary elements, then we just endow $Gal(T)$ with the pointwise convergence topology (that is we consider the set of imaginary elements with the discrete topology). Otherwise, it is more complicated, and here is what should be done in general:

For each $n \leq \omega$ and E invariant bounded equivalence relation on M^n, consider the canonical mapping φ_E from M^n onto M^n/E. By definition, a subset X of M^n/E is closed if and only if $\varphi_E^{-1}(X)$ is the intersection of a family of subsets definable with parameters. $Gal(T)$ acts on M^n/E and the topology on $Gal(T)$ is defined as the coarsest topology which makes all these actions (with various n and E) continuous.

Now, we can prove:

Theorem 5 *1. $Gal(T)$ is a topological compact group.*

2. *It is Hausdorff if and only if it acts faithfully on the set of bounded hyperimaginary elements, if and only if it acts faithfully on the set of finitary bounded hyperimaginary elements.*
3. *It is profinite if and only if it acts faithfully on the set of bounded imaginary elements.*

There is a Galois correspondence between the subgroups of $Gal(T)$ and the bounded ultra-imaginary elements: every subgroup of $Gal(T)$ is the stabilizer of an ultra-imaginary element. The hyperimaginary elements correspond to the closed subgroups and the imaginary elements correspond to the clopen subgroups of $Gal(T)$.

Let H_0 be the topological closure of the identity. Then H_0 is a normal subgroup of $Gal(T)$. If we consider $Gal(T)$ as a permutation group on $Bdd(T)$, H_0 is exactly the pointwise stabilizer of the set of bounded hyperimaginary elements. So, if we set $Gal_0(T) = Gal(T)/H_0$, $Gal_0(T)$ acts faithfully on the set of bounded hyperimaginary elements. As a quotient group, $Gal_0(T)$ is canonically endowed with a topology. This way, we get a compact Hausdorff group.

Recently, L. Newelski ([22]) has proved that H_0 is either trivial or of cardinality $2^{\aleph_0}$.

I would like to conclude this section by a conjecture. In all the known examples of countable saturated structures where the small index property is false, there is a

non open subgroup of $Gal(T)$ of countable index (and, its preimage by the canonical homomorphism from $Aut(M)$ onto $Gal(T)$ is a non open subgroup of $Aut(M)$ of countable index). If A has a cardinality strictly less than $card(M)$, define $Autf_A(M)$ as the subgroup of $Aut_A(M)$ generated by

$$\bigcup \{Aut_N(M) \ ; \ A \subseteq N \prec M\} .$$

The following conjecture is open, even in the ω-categorical case:

Conjecture 6 *Assume that M is a countable saturated structure and let H be a subgroup of $Autf(M)$ of index strictly less than $2^{\aleph_0}$. Then, there exists a finite subset $A \subset M$ such that $Autf_A(M) \subseteq H$.*

In [16], this conjecture is proved for almost strongly minimal sets (so, in particular for algebraically closed fields).

5. The uncountable case

We are now given a saturated structure M of cardinality $\lambda > \aleph_0$. The small index property has a natural generalization. If we assume that $\lambda^{<\lambda} = \lambda$ (i.e. there is exactly λ subsets of M of cardinality less than λ) then any subgroup of $Aut(M)$ containing a subgroup of the form $Aut_A(M)$ with $card(A) < \lambda$ has index at most λ. The converse is true:

Theorem 7 [20] *Assume that M is a saturated structure of uncountable cardinality $\lambda = \lambda^{<\lambda}$, and let H be a subgroup of $Aut(M)$ of cardinality at most λ. Then, there exists a subset A of M of cardinality less than λ such that $Aut_A(M) \subseteq H$.*

Here again, we may introduce a topological structure on $Aut(M)$: if μ is an infinite cardinal, let $\mathcal{T}_\mu$ be the group topology on $Aut(M)$ for which a basis of open neighborhoods of the identity is

$$\{Aut_A(M) \ ; \ A \subseteq M \text{ and } card(A) < \mu\} .$$

To complete this definition, let $\mathcal{T}_a(M)$ be the group topology on $Aut(M)$ for which a basis of open neighborhoods of the unit is

$$\{Autf_A(M) \ ; \ A \subseteq M \text{ and } A \text{ finite}\} .$$

The above theorem just says that the subgroups of $Aut(M)$ of index at most λ are exactly the open subgroups for $\mathcal{T}_\lambda(M)$, and consequently, the topology $\mathcal{T}_\lambda(M)$ can be reconstructed from the pure group structure. It is also clear that the open subsets for $\mathcal{T}_\lambda$ are just the intersections of less than λ $\mathcal{T}_a$-open subsets. So, if one knows $\mathcal{T}_a(M)$, one knows $\mathcal{T}_\lambda(M)$.

With a few cardinality hypotheses, we can reconstruct one topological group from another: (see [15] for details):

1. Let M and M' be two saturated models of the same theory. Then we can reconstruct $(Aut(M'), \mathcal{T}_a)$ from $(Aut(M), \mathcal{T}_a)$.
2. Let M and M' be two models of the same theory, and assume $card(M) = \lambda < \mu = card(M')$. Then we can reconstruct $Aut(M')$ from $(Aut(M), \mathcal{T}_\lambda)$ (and from $Aut(M)$ alone if $\lambda = \lambda^{<\lambda}$). In fact we can reconstruct $(Aut(M'), \mathcal{T}_\upsilon(M'))$ for every cardinal υ, $\lambda \leq \upsilon \leq \mu$.
3. Let M be a saturated structure of uncountable cardinality $\lambda = \mu^+ = 2^\mu$ and assume that T has a saturated model of cardinality μ. Then $(Aut(M), \mathcal{T}_a)$ can be reconstructed from $Aut(M)$.

Let us give an example of a theorem which can be proved using the above facts: Assume GCH and let T_1 and T_2 be two theories with saturated models M_1 and M_2 of cardinality μ^{++}. Assume that $Aut(M_1)$ and $Aut(M_2)$ are isomorphic. Then, for all cardinal λ, if T_1 has a saturated model of cardinality λ, then T_2 has also a saturated model in cardinality λ, and the automorphism groups of these two models are isomorphic.

References

[1] G. Ahlbrandt & M. Ziegler, Quasi finitely axiomatizable totally categorical theories, *Ann. Pure Appl. Logic* 30 (1986), 63-82.
[2] R. M. Bryant & D. M. Evans,. The small index property for free groups and relatively free group, *J. London Math. Soc.* (2) 55 (1997), no. 2, 363–369.
[3] S. Buechler, Lascar strong types in some simple theories, *J. Symbolic Logic* 64 (1999), no. 2, 817–824.
[4] E. Casanovas, D. Lascar, A. Pillay & M. Ziegler, Galois groups of first order theories, *J. Mathematical Logic* 1 no. 2, (2001), 305–311.
[5] J. D. Dixon, P. M. Neumann & S. Thomas, Subgroup of small index in infinite symmetric groups, *Bull. London Math. Soc.* 18 (1986), 580–586.
[6] D. M. Evans, Subgroups of small index in general linear groups, *Bull. London Math. Soc.* 18 (1986), 587–590.
[7] D. M. Evans & P. R. Hewitt, Counterexamples to a conjecture on relative categoricity, *Ann. Pure Appl. Logic* 46 no. 2 (1990), 201–209.
[8] B. Herwig, Extending partial automorphisms on finite structures, *Combinatoria* 15 (1995), 365–371.
[9] B. Herwig, Extending partial isomorphisms for the small index property of many ω-categorical structures, *Israel J. Math.* 107 (1998), 93–123.
[10] W. Hodges, I. Hodkinson, D. Lascar & S. Shelah, The small index property for ω-categorical ω-stable structures and for the random graph, *J. London Math. Soc.* (2) 48 (1993), 204–218.
[11] R. Kaye & H. Kotlarski, Automorphisms of models of true arithmetic: recognazing some open subgroups, *Notre Dame J. Forma Logic*, 2, 35 (1994), 1–14.
[12] B. Kim, A note on Lascar strong types in simple theories, *J. Symbolic Logic*, 63 no. 3 (1998), 926–936.
[13] R. Kossak & J. H. Schmerl, The automorphism group of an arithmetically

saturated model of Peano arithmetic, *J. London Math. Soc.* (2) 52 (1995), no. 2, 235–244.

[14] D. Lascar, On the category of models of a complete theory, *J. Symbolic Logic* 47 (1982), 249–266.

[15] D. Lascar, Autour de la propriété du petit indice, *Proc. London Math. Soc.* (3) 62 no.1 (1991), 25–53.

[16] D. Lascar, Les automorphismes d'un ensemble fortement minimal *J. Symbolic Logic* 57 no. 1 (1992) , 238–251.

[17] D. Lascar, The small index property and recursively saturated models of Peano arithmetic, *Automorphisms of first-order structures,* Oxford Sci. Publ., Oxford Univ. Press, New York, 1994, 281–292.

[18] D. Lascar, Recovering the action of an automorphism group, *Logic: from Foundations to Application, European logic Colloquium,* edited by W. Hodges and all, Clarendon Press (1996), 313–326.

[19] D. Lascar & A. Pillay, Hyperimaginaries and automorphism groups, *J. Symbolic Logic* 66, no. 1 (2001), 127-143.

[20] D. Lascar & S. Shelah, Uncountable saturated structures have the small index property, *Bull. London Math. Soc.* 25, no. 2 (1993), 125–131.

[21] H. D. Macpherson & P. M. Neumann, Subgroups of infinite symmetric groups, *J. London Math. Soc.* (2) 42 (1990), no. 1, 64–84.

[22] L. Newelski, The diameter of a Lascar strong type, *preprint*, (2002).

[23] S. W. Semmes, Endomorphisms of infinite symmetric groups, *Abstracts of the Am. Math. Soc.* 2 (1981), 426.

[24] M. Rubin, On the reconstruction of $\aleph_0$-categorical structures from their automorphism groups, *Proc. London Math. Soc.* (3) 69 no. 2 (1994), 225–249.

[25] J. K. Truss, Infinite permutation group; subgroups of small index, *J. Algebra* 120 (1989), 494–515.

Section 2. Algebra

S. Bigelow: *Representations of Braid Groups* 37

A. Bondal, D. Orlov: *Derived Categories of Coherent Sheaves* 47

M. Levine: *Algebraic Cobordism* 57

Cheryl E. Praeger: *Permutation Groups and Normal Subgroups* 67

Markus Rost: *Norm Varieties and Algebraic Cobordism* 77

Z. Sela: *Diophantine Geometry over Groups and the Elementary Theory of Free and Hyperbolic Groups* 87

J. T. Stafford: *Noncommutative Projective Geometry* 93

Dimitri Tamarkin: *Deformations of Chiral Algebras* 105

ICM 2002 · Vol. II · 37–45

Representations of Braid Groups

S. Bigelow*

Abstract

In this paper we survey some work on representations of B_n given by the induced action on a homology module of some space. One of these, called the Lawrence-Krammer representation, recently came to prominence when it was shown to be faithful for all n. We will outline the methods used, applying them to a closely related representation for which the proof is slightly easier. The main tool is the Blanchfield pairing, a sesquilinear pairing between elements of relative homology. We discuss two other applications of the Blanchfield pairing, namely a proof that the Burau representation is not faithful for large n, and a homological definition of the Jones polynomial. Finally, we discuss possible applications to the representation theory of the Hecke algebra, and ultimately of the symmetric group over fields of non-zero characteristic.

2000 Mathematics Subject Classification: 20F36, 20C08.
Keywords and Phrases: Braid groups, Configuration spaces, Homological representations, Blanchfield pairing.

1. Introduction

Artin's braid group B_n was originally defined as a group of geometric braids in $\mathbf{R}^3$. Representations of B_n have been studied for their own intrinsic interest, and also in connection to other areas of mathematics, most notably to knot invariants such as the Jones polynomial.

We will use the definition of B_n as the mapping class group of an n-times punctured disk D_n. A rich source of representations of B_n is the induced action on homology modules of spaces related to D_n. The Burau representation, one of the simplest and best known representations of braid groups, is most naturally defined as the induced action of B_n on the first homology of a cyclic covering space of D_n. Lawrence [9] extended this idea to configuration spaces in D_n, and was able to obtain all of the so-called Temperley-Lieb representations.

*Department of Mathematics & Statistics, University of Melbourne, Victoria 3010, Australia.
E-mail: bigelow@unimelb.edu.au

Lawrence's work seems to have received very little attention until one of her homological representations was shown to be faithful, thus proving that braid groups are linear and solving a longstanding open problem. Two independent and very different proofs of this have appeared in [1] and [8]. In this paper we will outline the former, emphasising the importance of the Blanchfield pairing. We then discuss two other applications of the Blanchfield pairing, namely the proof that the Burau representation is not faithful for large n, and a homological definition of the Jones polynomial of a knot. We conclude with some speculation on possible future applications to representations of Hecke algebras when q is a root of unity. These are related to representations of the symmetric group S_n over fields of bad characteristic, that is, fields in which $n! = 0$.

2. The Lawrence-Krammer representation

Let D be the unit disk centred at the origin in the complex plane. Fix arbitrary real numbers $-1 < p_1 < \cdots < p_n < 1$, which we will call "puncture points". Let

$$D_n = D \setminus \{p_1, \ldots, p_n\}$$

be the n-times punctured disk. The braid group B_n is the mapping class group of D_n, that is, the set of homeomorphisms from D_n to itself that act as the identity on ∂D, taken up to isotopy relative to ∂D. Let C be the space of all unordered pairs of distinct points in D_n.

Suppose x is a point in D_n, and α is a simple closed curve in D_n enclosing one puncture point and not enclosing x. Let $\gamma\colon I \to C$ be the loop in C given by

$$\gamma(s) = \{x, \alpha(s)\}.$$

Now suppose τ_1 and τ_2 are paths in D_n such that $\tau_1\tau_2$ is a simple closed curve that does not enclose any puncture points. Let $\tau\colon I \to C$ be the loop in C given by

$$\tau(s) = \{\tau_1(s), \tau_2(s)\}.$$

Let

$$\Phi\colon \pi_1(C) \to \langle q \rangle \oplus \langle t \rangle$$

be the unique homomorphism such that $\Phi(\gamma) = q$ and $\Phi(\tau) = t$ for any γ and τ defined as above. For a proof of the existence and uniqueness of such a homomorphism, see [11]. Let $\tilde{C}$ be the connected covering space of C whose fundamental group is the kernel of Φ.

The second homology $H_2(\tilde{C})$ is a module over $\mathbf{Z}[q^{\pm 1}, t^{\pm 1}]$, where q and t act by covering transformations. The Lawrence-Krammer representation of B_n is the induced action of B_n on $H_2(\tilde{C})$ by $\mathbf{Z}[q^{\pm 1}, t^{\pm 1}]$-module automorphisms. More precisely, given an element of B_n represented by a homeomorphism $\sigma\colon D_n \to D_n$, consider the induced action $\sigma\colon C \to C$. There is a unique lift $\tilde{\sigma}\colon \tilde{C} \to \tilde{C}$ that acts as the identity on $\partial\tilde{C}$. This induces an automorphism of $H_2(\tilde{C})$, which can be shown to respect the $\mathbf{Z}[q^{\pm 1}, t^{\pm 1}]$-module structure. See [11] for details.

3. The Blanchfield pairing

Let $\epsilon > 0$ be small. We define $P, B \subset C$ as follows. Suppose $\{x, y\}$ is a point in C. We say $\{x, y\} \in P$ if either $|x - y| \leq \epsilon$, or there is a puncture point p_i such that $|x - p_i| \leq \epsilon$ or $|y - p_i| \leq \epsilon$. We say $\{x, y\} \in B$ if $x \in \partial D$ or $y \in \partial D$.

For $u \in H_2(\tilde{C}, \tilde{P})$ and $v \in H_2(\tilde{C}, \tilde{B})$ let $(u \cdot v) \in \mathbf{Z}$ denote the standard algebraic intersection number. We define an intersection pairing

$$\langle \cdot, \cdot \rangle \colon H_2(\tilde{C}, \tilde{P}) \times H_2(\tilde{C}, \tilde{B}) \to \mathbf{Z}[q^{\pm 1}, t^{\pm 1}]$$

by

$$\langle u, v \rangle = \sum_{i,j \in \mathbf{Z}} (u \cdot q^i t^j v) q^i t^j.$$

For a proof that these are well-defined, see [7, Appendix E], where the following properties are also proved.

For $u \in H_2(\tilde{C}, \tilde{P})$, $v \in H_2(\tilde{C}, \tilde{B})$, $\sigma \in B_n$, and $\lambda \in \mathbf{Z}[q^{\pm 1}, t^{\pm 1}]$, we have

$$\langle \sigma u, \sigma v \rangle = \langle u, v \rangle,$$

and

$$\langle \lambda u, v \rangle = \lambda \langle u, v \rangle = \langle u, \bar{\lambda} v \rangle,$$

where $\bar{\lambda}$ is the image of λ under the automorphism of $\mathbf{Z}[q^{\pm 1}, t^{\pm 1}]$ taking q to q^{-1} and t to t^{-1}.

4. A faithful representation

The aim of this section is to outline a proof of the following.

Theorem. *Let $\tilde{C}$ and $\tilde{P}$ be as above. The induced action of B_n on $H_2(\tilde{C}, \tilde{P})$ is faithful.*

For the details, the reader is referred to [1], where the same techniques are used to prove that B_n acts faithfully on $H_2(\tilde{C})$. Our use of relative homology here actually simplifies the argument somewhat.

There is a slight technical difficulty in defining the action of B_n on $H_2(\tilde{C}, \tilde{P})$. Namely, the action of a braid on C need not preserve the set P. Thus we should really take a limit as ϵ approaches 0. The representation obtained is very similar to the Lawrence-Krammer representation, but has a slightly different module structure, as discussed in [3].

Let E be the straight edge from p_1 to p_2. Let E' be the set of points in C of the form $\{x, y\}$, where $x, y \in E$. Let $\tilde{E}'$ be a lift of E' to $\tilde{C}$. This represents an element of $H_2(\tilde{C}, \tilde{P})$, which we will call e. Let F_1 and F_2 be parallel vertical edges with endpoints on ∂D, passing between p_2 and p_3. Let F' be the set of points in C of the form $\{x, y\}$, where $x \in F_1$ and $y \in F_2$. Let $\tilde{F}'$ be a lift of F' to $\tilde{C}$. This represents an element of $H_2(\tilde{C}, \tilde{B})$, which we will call f. Note that

$$\langle e, f \rangle = 0,$$

since E' and F' are disjoint in C.

Suppose the action of B_n on $H_2(\tilde{C}, \tilde{P})$ is not faithful. It is not hard to show that there must be a braid σ in the kernel of this representation such that $\sigma(E)$ is not isotopic to E relative to endpoints. Now $\sigma(e) = e$, so

$$\langle \sigma(e), f \rangle = 0.$$

From this, we will derive a contradiction.

By applying an isotopy, we can assume $\sigma(E)$ intersects F_1 and F_2 transversely at a minimal number of points $k > 0$. Let $x_1, \ldots, x_k$ be the points of $\sigma(E) \cap F_1$, and let $y_1, \ldots, y_k$ be the points of $\sigma(E) \cap F_2$, numbered from top to bottom in both cases.

For $i, j \in \{1, \ldots, k\}$, let $a_{i,j}$ and $b_{i,j}$ be the unique integers such that $\sigma(\tilde{E}')$ intersects $q^{a_{i,j}} t^{b_{i,j}} \tilde{F}'$ at a point in the fibre over $\{x_i, y_j\}$, and let $\epsilon_{i,j}$ be the sign of that intersection. Then

$$\langle \sigma(e), f \rangle = \sum_{i=1}^{k} \sum_{j=1}^{k} \epsilon_{i,j} q^{a_{i,j}} t^{b_{i,j}}.$$

To calculate $a_{i,j}$ and $b_{i,j}$, it is necessary to specify choices of lift for E' and F'. We will not do this since we only need to calculate differences $a_{i',j'} - a_{i,j}$ and $b_{i',j'} - b_{i,j}$. To do this, let γ be a path in C that goes from $\{x_i, y_j\}$ to $\{x_{i'}, y_{j'}\}$ in $\sigma(E')$, and then back to $\{x_i, y_j\}$ in F'. Then

$$q^{(a_{i',j'} - a_{i,j})} t^{(b_{i',j'} - b_{i,j})} = \Phi(\gamma).$$

From this we can prove the following.

Lemma. *For all $i, j \in \{1, \ldots, k\}$ we have*

- $a_{i,j} = \frac{1}{2}(a_{i,i} + a_{j,j})$,
- *if $b_{i,j} > b_{i,i}$ then $a_{i,j'} > a_{i,i}$ for some $j' = 1, \ldots, k$,*
- *if $b_{i,j} > b_{j,j}$ then $a_{i',j} > a_{j,j}$ for some $i' = 1, \ldots, k$.*

The first of these is [1, Lemma 2.1], and the second and third follow from the proof of [1, Claim 3.4]. We now sketch the proof of the second in the special case where y_i lies between x_i and y_j along $\sigma(E)$.

Let α be the path from y_i to y_j along $\sigma(E)$. Let β be the path from y_j to y_i along F_2. Then $b_{i,j} - b_{i,i}$ is two times the winding number of $\alpha\beta$ around x_i. In particular, this winding number is positive.

Let D_1 be the once punctured disk $D \setminus \{x_i\}$, and let $\tilde{D}_1$ be its universal cover. Let $\tilde{\alpha}\tilde{\beta}$ be a lift of $\alpha\beta$ to D_1. This is a path from a point in the fibre over y_i to a "higher" point in the fibre over y_i.

Let F_2^+ be the the segment of F_2 going from y_i upwards to ∂D. Let $\tilde{F}_2^+$ be the lift of F_2^+ to $\tilde{D}_1$ that has an endpoint at $\tilde{\alpha}(0)$. In order to reach a higher sheet in $\tilde{D}_1$, $\tilde{\alpha}$ must intersect $\tilde{F}_2^+$. Let $\tilde{\gamma}$ be the loop in $\tilde{D}_1$ that follows $\tilde{\alpha}$ to the first point of intersection with $\tilde{F}_2^+$, and then follows $\tilde{F}_2^+$ back to $\tilde{\alpha}(0)$.

Let γ be the projection of $\tilde{\gamma}$ to D_1. This travels along $\sigma(E)$ from y_i to some point $y_{j'} \in F_2$, then along F_2 back to y_i. Then $a_{i,j'} - a_{i,i}$ is the total winding

number of γ around the puncture points. We must show that this winding number is positive.

By construction, $\tilde{\gamma}$ is a simple closed curve in $\tilde{D}_1$. By the Jordan curve theorem, it must bound a disk $\tilde{B}$. Let B be the projection of $\tilde{B}$ to D_1. This is an immersed disk in D, whose boundary is γ. Note that $\tilde{\gamma}$ passes anticlockwise around $\tilde{B}$, since the puncture x_i lies to its right. Thus the total winding number of γ around the puncture points is equal to the total number of puncture points contained in B, counted with multiplicities.

It remains to show that B intersects at least one puncture point. Suppose not. Then B is an immersed disk in D_n. Using an "innermost disk" argument, one can find an embedded disk B' in D_n whose boundary consists of a subarc of $\sigma(E)$ and a subarc of F_2. Using B', one can isotope $\sigma(E)$ so as to have fewer points of intersection with F_2, thus contradicting our assumptions.

This completes the proof of the second part of the lemma in the case where y_i lies between x_i and y_j along $\sigma(E)$. The remaining case, where x_i lies between y_i and y_j, is only slightly trickier. The third part of the lemma is similar to the second. The first part of the lemma is much easier.

We now return to the proof of the theorem. Let a be the maximum of all $a_{i,j}$. Let b be the maximum of $\{b_{i,j} : a_{i,j} = a\}$. Suppose $i, j \in \{1, \ldots, k\}$ are such that $a_{i,j} = a$ and $b_{i,j} = b$, and also $i', j' \in \{1, \ldots, k\}$ are such that $a_{i',j'} = a$ and $b_{i',j'} = b$. I claim that $\epsilon_{i,j} = \epsilon_{i',j'}$. From this claim, it follows that all occurrences of $q^a t^b$ in the expression

$$\langle \sigma(e), f \rangle = \sum_{i=1}^{k} \sum_{j=1}^{k} \epsilon_{i,j} q^{a_{i,j}} t^{b_{i,j}}$$

occur with the same sign, so the coefficient of $q^a t^b$ is non-zero in $\langle \sigma(e), f \rangle$. This provides our desired contradiction, and completes the proof of the theorem. It remains to prove that $\epsilon_{i',j'} = \epsilon_{i,j}$.

Using the above lemma, it is not hard to show that $a_{i,i} = a_{j,j} = a$ and $b_{i,i} = b_{j,j} = b$. Similarly, $a_{i',i'} = a_{j',j'} = a$ and $b_{i',i'} = b_{j',j'} = b$. We will only need the equalities

$$b_{i,i} = b_{i,j} = b_{j,j} = b_{i',i'} = b_{i',j'} = b_{j',j'}.$$

In fact, we only need these modulo two.

Orient $\sigma(E)$ so that it crosses F_1 from left to right at x_i. Let γ be the path in C which goes from $\{x_i, y_i\}$ to $\{x_{i'}, y_{i'}\}$ in E' and then back to $\{x_i, y_i\}$ in F'. Now $b_{i',i'} - b_{i,i}$ is the exponent of t in $\Phi(\gamma)$. The fact that this is an even number means that the pair of points in D_n do not "switch places" when they go around this loop. Thus $\sigma(E)$ crosses F_1 from left to right at x'_i. By similar arguments,

- $\beta(E)$ intersects F_1 with the same sign at x_i and $x_{i'}$,
- $\beta(E)$ intersects F_2 with the same sign at y_j and $y_{j'}$,
- x_i occurs before y_j and $x_{i'}$ occurs before $y_{j'}$ with respect to the orientation of $\sigma(E)$.

It is now intuitively clear that E' must intersect F' with the same signs at $\{x_i, y_j\}$ and $\{x_{i'}, y_{j'}\}$. This can be proved rigorously by careful consideration of orientations

of these surfaces, as discussed in [1, Section 2.1]. It follows that $\epsilon_{i,j} = \epsilon_{i',j'}$, which completes the proof of the theorem.

5. The Burau representation

The proof that the Lawrence-Krammer representation is faithful basically reduces to proving that the Blanchfield pairing detects whether corresponding edges in the disk can be isotoped to be disjoint. A converse to this idea leads to a proof that the Burau representation is not faithful for large n.

The Burau representation can be defined by a similar but simpler construction to that of the Lawrence-Krammer representation. Let

$$\Phi\colon \pi_1(D_n) \to \langle q \rangle$$

be the homomorphism that sends each of the obvious generators to q. Let $\tilde{D}_n$ be the corresponding covering space. The Burau representation is the induced action of B_n on $H_1(\tilde{D}_n)$ by $\mathbf{Z}[q^{\pm 1}]$-module automorphisms.

Let P be an ϵ-neighbourhood of the puncture points, and let $\tilde{P}$ be the preimage of P in $\tilde{D}_n$. The Blanchfield pairing in this context is a sesquilinear pairing

$$\langle \cdot, \cdot \rangle\colon H_1(\tilde{D}_n, \tilde{P}) \times H_1(\tilde{D}_n, \partial \tilde{D}_n) \to \mathbf{Z}[q^{\pm 1}].$$

Let E be the straight edge from p_1 to p_2. Let $\tilde{E}$ be a lift of E to $\tilde{D}_n$. This represents an element of $H_1(\tilde{D}_n, \tilde{P})$, which we will call e. Let F be a vertical edge with endpoints on ∂D, passing between p_{n-1} and p_n. Let $\tilde{F}$ be a lift of F to $\tilde{D}_n$. This represents an element of $H_1(\tilde{D}_n, \partial \tilde{D}_n)$, which we will call f. The following is [2, Theorem 5.1].

Theorem. *Let E, e, F and f be as above. The Burau representation of B_n is unfaithful if and only if there exists $\sigma \in B_n$ such that $\langle \sigma(e), f \rangle = 0$, but $\sigma(E)$ is not isotopic relative to endpoints to an edge that is disjoint from F.*

Using this theorem, one can show that the Burau representation of B_n is not faithful simply by exhibiting the required edges $\sigma(E)$ and F. Such edges can be found by hand in the case $n = 6$. In the case $n = 5$, they can be found by a computer search, and then laboriously checked by hand. The case $n = 4$ seems to be beyond the reach of any known computer algorithm. This is the last open case, since the Burau representation is known to be faithful for $n \leq 3$.

6. The Jones polynomial

In this section, we use the Blanchfield pairing to give a homological definition of the Jones polynomial of a knot or link. The Jones polynomial was defined in [6] using certain algebraically defined representations of braid groups. No satisfactory geometric definition is known, but some insight might be offered by defining the representations homologically and using the Blanchfield pairing. This was the original motivation for Lawrence to study homological representations of braid groups.

A *geometric braid* $\sigma \in B_n$ is a collection of n disjoint edges in $\mathbf{C} \times \mathbf{R}$ with endpoints $\{1, \ldots, n\} \times \{0, 1\}$, such that each edge goes from $\mathbf{C} \times \{0\}$ to $\mathbf{C} \times \{1\}$ with a constantly increasing $\mathbf{R}$ component. The correspondence between geometric braids and elements of the mapping class group is described in [1], and in many other introductory expositions on braids. The *plat closure* of a geometric braid $\sigma \in B_{2n}$ is the knot or link obtained by using straight edges to connect $(2j-1, k)$ to $(2j, k)$ for each $j = 1, \ldots, n$ and $k = 0, 1$. Every knot or link can be obtained in this way.

Let C be the configuration space of unordered n-tuples of distinct points in D_{2n}. Let

$$\Phi\colon \pi_1(C) \to \langle q \rangle \oplus \langle t \rangle$$

be defined as in Section 2. Namely, if γ is any loop in which one of the n-tuple winds anticlockwise around a puncture point, and τ is any loop in which two of the n-tuple exchange places by an anticlockwise half twist, then $\Phi(\gamma) = q$ and $\Phi(\tau) = t$. Let $\tilde{C}$ be the covering space corresponding to Φ.

Define $P, B \subset C$ similarly to those of Section 3. The Blanchfield pairing is a sesquilinear pairing

$$\langle \cdot, \cdot \rangle\colon H_2(\tilde{C}, \tilde{P}) \times H_2(\tilde{C}, \tilde{B}) \to \mathbf{Z}[q^{\pm 1}, t^{\pm 1}].$$

For $k = 1, \ldots, n$, let F_k be the straight edge from p_{2k-1} to p_{2k}. Let F be the set of points in C of the form $\{x_1, \ldots, x_n\}$ where $x_i \in F_i$. Let $\tilde{F}$ be a lift of F to $\tilde{C}$. This represents an element of $H_n(\tilde{C}, \tilde{P})$, which we call f. For $k = 1, \ldots, n$, let $e_k\colon S^1 \to D_n$ be a figure-eight around p_{2k-1} and p_{2k} in a small neighbourhood of F_k. Let e be the map from the n-torus $(S^1)^n$ to C given by

$$e(s_1, \ldots, s_n) = \{e_1(s_1), \ldots, e_n(s_n)\}.$$

Let $\tilde{e}$ be a lift of e to $\tilde{C}$. This represents an element of $H_n(\tilde{C})$, and hence of $H_n(\tilde{C}, \tilde{B})$, which we also call e.

The main result of [4] is the following.

Theorem. *Let e and f be as above and suppose $\sigma \in B_{2n}$. The Jones polynomial of the plat closure of σ is*

$$\langle \sigma(e), f) \rangle|_{(t=-q^{-1})},$$

up to sign and multiplication by a power of $q^{\frac{1}{2}}$.

Here, the Jones polynomial is normalised so that the Jones polynomial of the unknot is $-q^{\frac{1}{2}} - q^{-\frac{1}{2}}$. The correct sign and power of $q^{\frac{1}{2}}$ is also explicitly specified in [4].

This result is due to Lawrence, who also achieved a similar result for the two-variable Jones polynomial by a much more complicated construction [10].

7. The Hecke algebra

We conclude with some speculation about possible applications of the Blanchfield pairing to the representation theory of Hecke algebras. We first give a very brief overview of the basic theory of Hecke algebras.

Let $q \in \mathbf{C} \setminus \{0\}$. The Hecke algebra $H_n(q)$, or simply H_n, is the $\mathbf{C}$-algebra given by generators $g_1, \ldots, g_{n-1}$ and relations

- $g_i g_j = g_j g_i$ if $|i-j| > 1$,
- $g_i g_j g_i = g_i g_j g_i$ if $|i-j| = 1$,
- $(g_i - 1)(g_i + q) = 0$.

It is an $n!$-dimensional $\mathbf{C}$-algebra. We are restricting to the ring $\mathbf{C}$ for convenience, although other rings can be used.

Note that $H_n(1)$ is the group algebra $\mathbf{C}S_n$ of the symmetric group S_n. The Hecke algebra is called a "quantum deformation" of $\mathbf{C}S_n$. The representation theory of $\mathbf{C}S_n$ is well understood except when working over a field of finite characteristic in which $n! = 0$. This is because the classical theory sometimes requires one to divide by $n!$, the order of the group. When studying H_n it turns out to be useful to be able to divide by

$$(1 + q + \cdots + q^{n-1})(1 + q + \cdots + q^{n-2}) \ldots (1 + q).$$

This is sometimes written as $[n]!$, and can be thought of as a "quantum deformation" of $n!$. Note that $[n]! = n!$ if $q = 1$. A *generic* value of q is one for which $[n]! \neq 0$. The non-generic values are the primitive kth roots of unity for $k = 2, \ldots, n$. The representation theory of H_n is well understood for generic values of q, but the non-generic values are the subject of ongoing active research.

One of the most important papers on this subject is Dipper and James [5]. For every partition λ of n, they define a H_n-module S^λ called the *Specht module.* They then define a bilinear pairing on S^λ, which we denote $\langle \cdot, \cdot \rangle_{\mathrm{DJ}}$. Let $S^\lambda_\perp$ denote the set of $u \in S^\lambda$ such that $\langle u, v \rangle_{\mathrm{DJ}} = 0$ for all $v \in S^\lambda$. Let D^λ be the quotient module $S^\lambda / S^\lambda_\perp$. Dipper and James show that the non-zero D^λ form a complete list of all distinct irreducible representations of H_n. For generic values of q we have $D^\lambda = S^\lambda$. For non-generic values of q, the D^λ are not well understood.

Lawrence [10] gave a homological definition of the Specht modules. The construction begins with the action of B_n on a homology module of a configuration space. The variable t is then specialised to $-q^{-1}$, and a certain quotient module is taken. A detailed treatment of the case $\lambda = (n-2, 2)$ is given in [3].

There is a Blanchfield pairing on the Specht modules as defined by Lawrence. It would be nice if this were the same as the pairing defined by Dipper and James. Unfortunately the Blanchfield pairing is sesquilinear, whereas the pairing defined by Dipper and James is bilinear. This problem can be overcome as follows. Let $\rho\colon D_n \to D_n$ be the conjugation map. Let $\tilde{\rho}$ be the induced map on $H_k(\tilde{C}, \tilde{B})$. Then the pairing

$$\langle u, v \rangle' = \langle u, \rho(v) \rangle$$

can be shown to give a bilinear pairing on the Specht module.

For generic values of q, this topologically defined pairing is the same as the algebraically defined pairing of Dipper and James, up to some renormalisation.

There is some evidence that this can be made to work at non-generic values of q. If so, it would give rise to a new homological definition of the modules D^λ, and new topological tools for studying them. In any case, it would be interesting to better understand the behaviour of this Blanchfield pairing at roots of unity.

References

[1] S. Bigelow, Braid groups are linear, *J. Amer. Math. Soc.*, 14 (2001), 471–486.

[2] S. Bigelow, Does the Jones polynomial detect the unknot? *J. Knot Theory Ramifications*, (to appear).

[3] S. Bigelow, The Lawrence-Krammer representation, *Proceedings, Georgia Topology Conference, 2001*, (to appear).

[4] S. Bigelow, A homological definition of the Jones polynomial, *Proceedings, RIMS, Kyoto, 2001*, (to appear).

[5] R. Dipper & G. James, Representations of Hecke algebras of general linear groups, *Proc. London Math. Soc. (3)*, 52 (1986), 20–52.

[6] V. Jones, A polynomial invariant for knots via von Neumann algebras, *Bull. Amer. Math. Soc. (N.S.)*, 12 (1985), 103–111.

[7] A. Kawauchi, *A survey of knot theory*, Birkhäuser Verlag, 1996.

[8] D. Krammer, Braid groups are linear, *Ann. of Math. (2)*, 155 (2002), 131–156.

[9] R. Lawrence, Homological representations of the Hecke algebra, *Comm. Math. Phys.*, 135 (1990), 141–191.

[10] R. Lawrence, Braid group representations associated with sl_m, *J. Knot Theory Ramifications*, 5 (1996), 637–660.

[11] L. Paoluzzi & L. Paris, A note on the Lawrence-Krammer-Bigelow representation, *Algebr. Geom. Topol.*, 2 (2002), 499–518.

ICM 2002 · Vol. II · 47–56

Derived Categories of Coherent Sheaves

A. Bondal* **D. Orlov†**

Abstract

We show how derived categories build bridges across the current mathematical mainstream, linking geometric and algebraic, commutative and noncommutative, local and global banks. Arches in these bridges are pieces of semiorthogonal decompositions of triangulated categories.

2000 Mathematics Subject Classification: 18E30, 14F05.
Keywords and Phrases: Derived categories, Coherent sheaves, Fully faithful functors, Noncommutative geometry.

1. Introduction

This paper is devoted to studying the derived categories $\mathcal{D}^b(X)$ of coherent sheaves on smooth algebraic varieties X and on their noncommutative counterparts. Derived categories of coherent sheaves proved to contain the complete geometric information about varieties (in the sense of the classical Italian school of algebraic geometry) as well as the related homological algebra.

The situation when there exists a functor $\mathcal{D}^b(M) \longrightarrow \mathcal{D}^b(X)$ which is fully faithful is of special interest. We are convinced that any example of such a functor is both algebraically and geometrically meaningful.

A particular case of a fully faithful functor is an equivalence of derived categories $\mathcal{D}^b(M) \xrightarrow{\sim} \mathcal{D}^b(X)$.

We show that smooth projective varieties with ample canonical or anticanonical bundles are uniquely determined by their derived categories. Hence the derived equivalences between them boil down to autoequivalences. We prove that for such a variety the group of exact autoequivalences is the semidirect product of the group of automorphisms of the variety and the Picard group plus translations.

Equivalences and autoequivalences for the case of varieties with non-ample (anti) canonical sheaf are now intensively studied. The group of autoequivalences

*Algebra Section, Steklov Mathematical Institute, Russian Academy of Sciences, 8 Gubkin St., GSP-1, Moscow 117966, Russia. E-mail: bondal@mi.ras.ru
†Algebra Section, Steklov Mathematical Institute, Russian Academy of Sciences, 8 Gubkin St., GSP-1, Moscow 117966, Russia. E-mail: orlov@mi.ras.ru

is believed to be closely related to the holonomy group of the mirror-symmetric family.

We give a criterion for a functor between derived categories of coherent sheaves on two algebraic varieties to be fully faithful. Roughly speaking, it is in orthogonality of the images under the functor of the structure sheaves of distinct closed points of the variety. If a functor $\Phi : \mathcal{D}^b(M) \longrightarrow \mathcal{D}^b(X)$ is fully faithful, then it induces a so-called semiorthogonal decomposition of $\mathcal{D}^b(X)$ into $\mathcal{D}^b(M)$ and its right orthogonal category.

It turned out that derived categories have nice behavior under special birational transformations like blow ups, flips and flops. We describe a semiorthogonal decomposition of the derived category of the blow-up of a smooth variety X in a smooth center $Y \subset X$. It contains one component isomorphic to $\mathcal{D}^b(X)$ and several components isomorphic to $\mathcal{D}^b(Y)$.

We also consider some flips and flops. Examples support the conjecture that for any generalized flip $X \dashrightarrow X^+$ there exists a fully faithful functor $\mathcal{D}^b(X^+) \to \mathcal{D}^b(X)$ and it must be an equivalence for generalized flops. This suggests the idea that the minimal model program of the birational geometry can be viewed as a 'minimization' of the derived category $\mathcal{D}^b(X)$ in a given birational class of X.

Then we widen the categorical approach to birational geometry by including in the picture some noncommutative varieties. We propose to consider noncommutative desingularizations and formulate a conjecture generalizing the derived McKay correspondence.

We construct a semiorthogonal decomposition for the derived category of the complete intersections of quadrics. It is related to classical questions of algebraic geometry, like 'quadratic complexes of lines', and to noncommutative geometric version of Koszul quadratic duality.

2. Equivalences between derived categories

The first question that arises in studying algebraic varieties from the point of view of derived categories is when varieties have equivalent derived categories of coherent sheaves. Examples of such equivalences for abelian varieties and K3 surfaces were constructed by Mukai [Mu1], [Mu2], Polishchuk [Po] and the second author in [Or2], [Or3]. See below on derived equivalences for birational maps.

Yet we prove that a variety X is uniquely determined by its category $\mathcal{D}^b(X)$, if its anticanonical (Fano case) or canonical (general type case) sheaf is ample. To this end, we use only the graded (not triangulated) structure of the category. By definition a graded category is a pair $(\mathcal{D}, T_{\mathcal{D}})$ consisting of a category $\mathcal{D}$ (which we always assume to be k-linear over a field k) and a fixed equivalence $T_D : \mathcal{D} \longrightarrow \mathcal{D}$, called translation functor. For derived categories the translation functor is defined to be the shift of grading in complexes.

Of crucial importance for exploring derived categories are existence and properties of the Serre functor, defined in [BK].

Definition 1 [BK] [BO2] *Let $\mathcal{D}$ be a k-linear category with finite-dimensional* Hom*'s. A covariant additive functor $S : \mathcal{D} \to \mathcal{D}$ is called a* Serre functor *if it*

is an equivalence and there are given bi–functorial isomorphisms for any $A, B \in \mathcal{D}$*:*

$$\varphi_{A,B} : \mathrm{Hom}_{\mathcal{D}}(A\ ,\ B) \xrightarrow{\sim} \mathrm{Hom}_{\mathcal{D}}(B\ ,\ SA)^*.$$

A Serre functor in a category $\mathcal{D}$, if it exists, is unique up to a graded natural isomorphism.

If X is a smooth algebraic variety, $n = \dim X$, then the functor $(\cdot) \otimes \omega_X[n]$ is the Serre functor in $\mathcal{D}^b(X)$. Thus, the Serre functor in $\mathcal{D}^b(X)$ can be viewed as a categorical incarnation of the canonical sheaf ω_X.

Theorem 2 [BO2] *Let* X *be a smooth irreducible projective variety with ample canonical or anticanonical sheaf. If* $\mathcal{D}^b(X)$ *is equivalent as a graded category to* $\mathcal{D}^b(X')$ *for some other smooth algebraic variety* X'*, then* X *is isomorphic to* X'*.*

The idea of the proof is that for varieties with ample canonical or anticanonical sheaf we can recognize the skyscraper sheaves of closed points in $\mathcal{D}^b(X)$ by means of the Serre functor. In this way we find the variety as a set. Then we reconstruct one by one the set of line bundles, Zariski topology and the structural sheaf of rings.

This theorem has a generalization to smooth orbifolds related to projective varieties with mild singularities, as it was shown by Y. Kawamata [Kaw].

Now consider the problem of computing the group $\mathrm{Aut}\mathcal{D}^b(X)$ of exact (i.e. preserving triangulated structure) autoequivalences of $\mathcal{D}^b(X)$ for an individual X.

Theorem 3 [BO2] *Let* X *be a smooth irreducible projective variety with ample canonical or anticanonical sheaf. Then the group of isomorphism classes of exact autoequivalences* $\mathcal{D}^b(X) \to \mathcal{D}^b(X)$ *is generated by the automorphisms of the variety, twists by all invertible sheaves and translations.*

In the hypothesis of Theorem 3 the group $\mathrm{Aut}\mathcal{D}^b(X)$ is the semi-direct product of its subgroups $\mathrm{Pic}X \oplus \mathbf{Z}$ and $\mathrm{Aut}X$, $\mathbf{Z}$ being generated by the translation functor:

$$\mathrm{Aut}\mathcal{D}^b(X) \cong \mathrm{Aut}X \ltimes (\mathrm{Pic}X \oplus \mathbf{Z}).$$

3. Fully faithful functors and semiorthogonal decompositions

An equivalence is a particular instance of a fully faithful functor. This is a functor $F : \mathcal{C} \to \mathcal{D}$ which for any pair of objects $X, Y \in \mathcal{C}$ induces an isomorphism $\mathrm{Hom}(X\ ,\ Y) \simeq \mathrm{Hom}(FX\ ,\ FY)$. This notion is especially useful in the context of triangulated categories.

If a functor $\Phi : \mathcal{D}^b(M) \longrightarrow \mathcal{D}^b(X)$ is fully faithful, then it induces a so-called semiorthogonal decomposition of $\mathcal{D}^b(X)$ into $\mathcal{D}^b(M)$ and its right orthogonal.

Let $\mathcal{B}$ be a full triangulated subcategory of a triangulated category $\mathcal{D}$. The right orthogonal to $\mathcal{B}$ is the full subcategory $\mathcal{B}^{\perp} \subset \mathcal{D}$ consisting of the objects C such that $\mathrm{Hom}(B\ ,\ C) = 0$ for all $B \in \mathcal{B}$. The left orthogonal ${}^{\perp}\mathcal{B}$ is defined analogously. The categories ${}^{\perp}\mathcal{B}$ and $\mathcal{B}^{\perp}$ are also triangulated.

Definition 4 [BK] *A sequence of triangulated subcategories* $(\mathcal{B}_0, ..., \mathcal{B}_n)$ *in a triangulated category* $\mathcal{D}$ *is said to be* semiorthogonal *if* $\mathcal{B}_j \subset \mathcal{B}_i^{\perp}$ *whenever* $0 \leq j < i \leq n$. *If a semiorthogonal sequence generates* $\mathcal{D}$ *as a triangulated category, then we call it by* semiorthogonal decomposition *of the category* $\mathcal{D}$ *and denote this as follows:*

$$\mathcal{D} = \left\langle \mathcal{B}_0, ..., \mathcal{B}_n \right\rangle.$$

Examples of semiorthogonal decompositions are provided by exceptional sequences of objects [Bo]. These arise when all $\mathcal{B}_i$'s are equivalent to the derived categories of finite dimensional vector spaces $\mathcal{D}^b(k - \mathrm{mod})$. Objects which correspond to the 1-dimensional vector space under a fully faithful functor $F : \mathcal{D}^b(k - \mathrm{mod}) \to \mathcal{D}$ can be intrinsically defined as *exceptional*, i.e. those satisfying the conditions $\mathrm{Hom}^i(E\ ,\ E) = 0$, when $i \neq 0$, and $\mathrm{Hom}^0(E\ ,\ E) = k$. There is a natural action of the braid group on exceptional sequences [Bo] and, under some conditions, on semiorthogonal sequences of subcategories in a triangulated category [BK].

We propose to consider the derived category of coherent sheaves as an analogue of the motive of a variety, and semiorthogonal decompositions as a tool for simplification of this 'motive' similar to splitting by projectors in Grothendieck motivic theory.

Let X and M be smooth algebraic varieties of dimension n and m respectively and E an object in $\mathcal{D}^b(X \times M)$. Denote by p and π the projections of $M \times X$ to M and X respectively. With E one can associate the functor $\Phi_E : \mathcal{D}^b(M) \longrightarrow \mathcal{D}^b(X)$ defined by the formula:

$$\Phi_E(\cdot) := \mathbf{R}\pi_*(E \overset{\mathbf{L}}{\otimes} p^*(\cdot)).$$

It happens that any fully faithful functor is of this form.

Theorem 5 [Or2] *Let* $F : \mathcal{D}^b(M) \to \mathcal{D}^b(X)$ *be an exact fully faithful functor, where* M *and* X *are smooth projective varieties. Then there exists a unique up to isomorphism object* $E \in \mathcal{D}^b(M \times X)$ *such that* F *is isomorphic to the functor* Φ_E.

The assumption on existence of the right adjoint to F, which was originally in [Or2], can be removed in view of saturatedness of $\mathcal{D}^b(M)$ due to [BK], [BVdB].

This theorem is in conjunction with the following criterion.

Theorem 6 [BO1] *Let* M *and* X *be smooth algebraic varieties and* $E \in \mathcal{D}^b(M \times X)$. *Then* Φ_E *is fully faithful functor if and only if the following orthogonality conditions are verified:*

$i)$ $\mathrm{Hom}^i_X(\Phi_E(\mathcal{O}_{t_1})\ ,\ \Phi_E(\mathcal{O}_{t_2})) = 0$ *for every* i *and* $t_1 \neq t_2$.

$ii)$ $\mathrm{Hom}^0_X(\Phi_E(\mathcal{O}_t)\ ,\ \Phi_E(\mathcal{O}_t)) = k,$

$\mathrm{Hom}^i_X(\Phi_E(\mathcal{O}_t)\ ,\ \Phi_E(\mathcal{O}_t)) = 0,$ *for* $i \notin [0, dim M]$.

Here t, t_1, t_2 *stand for closed points in* M *and* $\mathcal{O}_{t_i}$ *for the skyscraper sheaves.*

The criterion is a particular manifestation of the following important principle: suppose M is realized as an appropriate moduli space of pairwise homologically orthogonal objects in a triangulated category $\mathcal{D}$ taken 'from real life', then one can expect a sheaf of finite (noncommutative) algebras $\mathcal{A}_M$ over $\mathcal{O}_M$ and a fully faithful functor from the derived category $\mathcal{D}^b(coh(\mathcal{A}_M))$ of coherent modules over $\mathcal{A}_M$ to $\mathcal{D}$.

There are also strong indications that this principle should have a generalization, at the price of considering noncommutative DG moduli spaces, to the case when the orthogonality condition is dropped.

4. Derived categories and birational geometry

Behavior of derived categories under birational transformations shows that they can serve as a useful tool in comprehending various phenomena of birational geometry and play possibly the key role in realizing the minimal model program.

First, we give a description of the derived category of the blow-up of a variety with smooth center in terms of the categories of the variety and of the center. Let Y be a smooth subvariety of codimension r in a smooth algebraic variety X. Denote $\widetilde{X}$ the smooth algebraic variety obtained by blowing up X along the center Y. There exists a fibred square:

$$\begin{array}{ccc} \widetilde{Y} & \stackrel{j}{\longrightarrow} & \widetilde{X} \\ p\downarrow & & \pi\downarrow \\ Y & \stackrel{i}{\longrightarrow} & X \end{array}$$

where i and j are smooth embeddings, and $p : \widetilde{Y} \to Y$ is the projective fibration of the exceptional divisor $\widetilde{Y}$ in $\widetilde{X}$ over the center Y. Recall that $\widetilde{Y} \cong \mathbb{P}(N_{X/Y})$ is the projective normal bundle. Denote by $\mathcal{O}_{\widetilde{Y}}(1)$ the relative Grothendieck sheaf.

Proposition 7 [Or1] *Let L be any invertible sheaf on $\widetilde{Y}$. The functors*

$$\mathbf{L}\pi^* : \mathcal{D}^b(X) \longrightarrow \mathcal{D}^b(\widetilde{X}),$$

$$\mathbf{R}j_*(L \otimes p^*(\cdot)) : \mathcal{D}^b(Y) \longrightarrow \mathcal{D}^b(\widetilde{X})$$

are fully faithful.

Denote by $D(X)$ the full subcategory of $\mathcal{D}^b(\widetilde{X})$ which is the image of $\mathcal{D}^b(X)$ with respect to the functor $\mathbf{L}\pi^*$ and by $D(Y)_k$ the full subcategories of $\mathcal{D}^b(\widetilde{X})$ which are the images of $\mathcal{D}^b(Y)$ with respect to the functors $\mathbf{R}j_*(\mathcal{O}_{\widetilde{Y}}(k) \otimes p^*(\cdot))$.

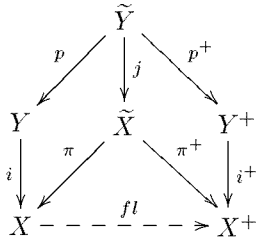
Theorem 8 [Or1][BO1] *We have the semiorthogonal decomposition of the category of the blow-up:*

$$\mathcal{D}^b(\widetilde{X}) = \Big\langle D(Y)_{-r+1}, ..., D(Y)_{-1}, D(X) \Big\rangle.$$

Now we consider the behavior of the derived categories of coherent sheaves with respect to the special birational transformations called flips and flops.

Let Y be a smooth subvariety of a smooth algebraic variety X such that $Y \cong \mathbb{P}^k$ and $N_{X/Y} \cong \mathcal{O}(-1)^{\oplus(l+1)}$ with $l \leq k$.

If $\widetilde{X}$ is the blow-up of X along Y, then the exceptional divisor $\widetilde{Y} \cong \mathbb{P}^k \times \mathbb{P}^l$ is the product of projective spaces. We can blow down $\widetilde{X}$ in such a way that $\widetilde{Y}$ projects to the second component $\mathbb{P}^l$ of the product. As a result we obtain a smooth variety X^+, which for simplicity we assume to be algebraic, with subvariety $Y^+ \cong \mathbb{P}^l$. This is depicted in the following diagram:

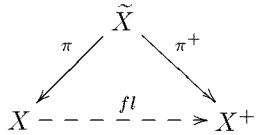


The birational map $X \dashrightarrow X^+$ is the simplest instance of flip, for $l \leq k$. If $l = k$, this is a flop.

Theorem 9 [BO1] *In the above notations, the functor* $\mathbf{R}\pi_*\mathbf{L}\pi^{+*} : \mathcal{D}^b(X^+) \longrightarrow \mathcal{D}^b(X)$ *is fully faithful for* $l \leq k$. *If* $l = k$, *it is an equivalence.*

This theorem has an obvious generalization to the case when Y is isomorphic to the projectivization of a vector bundle E of rank k on a smooth variety W, $q : Y \longrightarrow W$, and $N_{X/Y} = q^*F \otimes \mathcal{O}_E(-1)$ where F is a vector bundle on W of rank $l \leq k$. Then the blow-up with a smooth center can be viewed as the particular case of this flip when Y is a divisor in X. Kawamata [Kaw] generalized the theorem to those flips between smooth orbifolds which are elementary (Morse type) cobordisms in the theory of birational cobordisms due to Wlodarczyk *et al.* [Wl], [AKMW].

Let X and X^+ be smooth projective varieties. A birational map $X \overset{fl}{\dashrightarrow} X^+$ will be called a generalized flip if for some (and consequently for any) its smooth resolution



the difference $D = \pi^*K_X - \pi^{+*}K_{X^+}$ between the pull-backs of the canonical divisors is an effective divisor on $\widetilde{X}$. The particular case when $D = 0$ is called generalized flop.

Theorem 9 together with calculations of 3-dimensional flops with centers in (-2)-curves [BO1] lead us to the following conjecture.

Conjecture 10 *For any generalized flip* $X \overset{fl}{\dashrightarrow} X^+$ *there is an exact fully faithful functor* $F : \mathcal{D}^b(X^+) \longrightarrow \mathcal{D}^b(X)$. *It is an equivalence for generalized flops.*

This conjecture was recently proved in dimension 3 by T. Bridgeland [Br].

The functor $\mathbf{R}\pi_*\mathbf{L}\pi^{+*} : \mathcal{D}^b(X^+) \longrightarrow \mathcal{D}^b(X)$ is not always fully faithful under conditions of the conjecture, but we expect that it is such when $\widetilde{X}$ is replaced by the fibred product of X and X^+ over some common singular contraction of X and X^+. Namikawa proved that this is the case for Mukai symplectic flops [Na].

A fully faithful functor $\mathcal{D}^b(X^+) \longrightarrow \mathcal{D}^b(X)$ induces a semiorthogonal decomposition of $\mathcal{D}^b(X)$ into $\mathcal{D}^b(X^+)$ and its right orthogonal (which is trivial for flops). Hence, passing from X to X^+ for generalized flips has the categorical meaning of breaking off semiorthogonal summands from the derived category. This suggests the idea that the minimal model program of birational geometry should be interpreted as a minimization of the derived category $\mathcal{D}^b(X)$ in a given birational class of X. Promisingly, chances are that the very existence of flips can be achieved by constructing X^+ as an appropriate moduli space of objects in $\mathcal{D}^b(X)$, in accordance with the principle of the previous section (this is done by T. Bridgeland for flops in dimension 3 [Br]).

5. Noncommutative resolutions of singularities

In this section we will give a perspective for categorical interpretation of the minimal model program by enriching the landscape with the derived categories of noncommutative varieties.

Let $\pi : \widetilde{X} \to X$ be a proper birational morphism, where X has rational singularities. Then $\mathbf{R}\pi_* : \mathcal{D}^b(\widetilde{X}) \to \mathcal{D}^b(X)$ identifies $\mathcal{D}^b(X)$ with the quotient of $\mathcal{D}^b(\widetilde{X})$ by the kernel of $\mathbf{R}\pi_*$. For this reason, let us call by a *categorical desingularization* of a triangulated category $\mathcal{D}$ a pair $(\mathcal{C}, \mathcal{K})$ consisting of an abelian category $\mathcal{C}$ of finite homological dimension and of $\mathcal{K}$, a thick subcategory in $\mathcal{D}^b(\mathcal{C})$ such that $\mathcal{D} = \mathcal{D}^b(\mathcal{C})/\mathcal{K}$. We expect that for $\mathcal{D} = \mathcal{D}^b(X)$ there exists a minimal desingularization, i.e. such one that $\mathcal{D}^b(\mathcal{C})$ has a fully faithful embedding in $\mathcal{D}^b(\mathcal{C}')$ for any other categorical desingularization $(\mathcal{C}', \mathcal{K}')$ of $\mathcal{D}$. Such a desingularization is unique up to derived equivalence of $\mathcal{C}$.

For the derived categories of singular varieties one may hope to find the minimal desingularizations in the spirit of noncommutative geometry.

Let X be an algebraic variety. We call by noncommutative (birational) desingularization of X a pair $(p, \mathcal{A})$ consisting of a proper birational morphism $p : Y \to X$ and an algebra $\mathcal{A} = \mathcal{E}nd(\mathcal{F})$ on Y, the sheaf of local endomorphisms of a torsion free coherent $\mathcal{O}_Y$-module $\mathcal{F}$, such that the abelian category of coherent $\mathcal{A}$-modules has finite homological dimension.

When $f : Y \to X$ is a morphism from a smooth Y onto an affine X with fibres of dimension ≤ 1 and $\mathbf{R}f_*(\mathcal{O}_Y) = \mathcal{O}_X$, M. Van den Bergh [VdB] has recently constructed a noncommutative desingularization of X, which is derived equivalent to $\mathcal{D}^b(Y)$.

Conjecture 11 *Let X has canonical singularities and $q : Y \to X$ a finite morphism with smooth Y. Then the pair $(id_X, \mathcal{E}nd(q_*\mathcal{O}_Y))$ is a minimal desingularization of X.*

In particular, we expect that $\mathcal{D}^b(coh(\mathcal{E}nd(q_*\mathcal{O}_Y)))$ has a fully faithful functor into $\mathcal{D}^b(\widetilde{X})$ for any (commutative) resolution of X. Moreover, if the resolution is crepant then the functor has to be an equivalence.

Let X be the quotient of a smooth Y by an action of a finite group G. If the locus of the points in Y with nontrivial stabilizer in G has codimension ≥ 2, then the category of coherent $\mathcal{E}nd(q_*\mathcal{O}_Y)$-modules is equivalent to the category of G-equivariant coherent sheaves on Y. Therefore the conjecture is a generalization of the derived McKay correspondence due to Bridgeland-King-Reid [BKR].

6. Complete intersection of quadrics and noncommutative geometry

This section is related to the previous one by Grothendieck slogan that projective geometry is a part of theory of singularities.

Let X be a smooth intersection of two projective quadrics of even dimension d over an algebraically closed field of characteristic zero. It appears that if we consider the hyperelliptic curve C which is the double cover of $\mathbb{P}^1$ that parameterizes the pencil of quadrics, with ramification in the points corresponding to degenerate quadrics, then $\mathcal{D}^b(C)$ is embedded in $\mathcal{D}^b(X)$ as a full subcategory [BO1]. This gives a categorical explanation for the classical description of moduli spaces of semistable bundles on the curve C as moduli spaces of (complexes of) coherent sheaves on X.

The orthogonal to $\mathcal{D}^b(C)$ in $\mathcal{D}^b(X)$ is decomposed into an exceptional sequence (of line bundles). More precisely, we have a semiorthogonal decomposition

$$\mathcal{D}^b(X) \;=\; \Big\langle \mathcal{O}_X(-d+3), \ldots, \mathcal{O}_X, \mathcal{D}^b(C) \Big\rangle. \tag{6.1}$$

When a greater number of quadrics is intersected, objects of noncommutative geometry naturally show up: instead of coherent sheaves on hyperelliptic curves we must consider modules over a sheaf of noncommutative algebras. More about noncommutative geometry is in the talk of T. Stafford at this Congress.

Consider a system of m quadrics in $\mathbb{P}(V)$, i.e. a linear embedding $U \overset{\phi}{\hookrightarrow} S^2V^*$, where $\dim U = m$, $\dim V = n$, $2m \leq n$. Let X, the complete intersection of the quadrics, be a smooth subvariety in $\mathbb{P}(V)$ of dimension $n-m-1$. Let $A = \bigoplus_{i\geq 0} H^0(X, \mathcal{O}(i))$ be the coordinate ring of X. This graded quadratic algebra is Koszul due to Tate [Ta]. The quadratic dual algebra $B = A^!$ is the generalized homogeneous Clifford algebra. It is generated in degree 1 by the space V, the relations being given by the kernel of the dual to ϕ map $S^2V \to U^*$, viewed as a subspace in $V \otimes V$. The center of B is generated by U^* (a subspace of quadratic elements in B) and an element d, which satisfies the equation $d^2 = f$ where f is the equation of the locus of degenerate quadrics in U. Algebra B is finite over the central subalgebra $S = S^\bullet U^*$. The Veronese subalgebra $B_{ev} = \oplus B_{2i}$ is finite over the Veronese subalgebra $S_{ev} = \oplus S^{2i}U^*$. Since **Proj** S_{ev} is isomorphic to $\mathbb{P}(U)$, the sheafification of B_{ev} over **Proj** S_{ev} is a sheaf $\mathcal{B}$ of finite algebras over $\mathcal{O}_{\mathbb{P}(U)}$. Consider the derived category $\mathcal{D}^b(coh(\mathcal{B}))$ of coherent right $\mathcal{B}$-modules.

Theorem 12 *Let X be the smooth intersection of m quadrics in $\mathbb{P}^{n-1}$, $2m \leq n$. Then there exists a fully faithful functor $\mathcal{D}^b(coh(\mathcal{B})) \to \mathcal{D}^b(X)$. Moreover,*

(i) *if $2m < n$, we have a semiorthogonal decomposition*

$$\mathcal{D}^b(X) = \Big\langle \mathcal{O}_X(-n+2m+1), \ldots, \mathcal{O}_X, \mathcal{D}^b(coh(\mathcal{B})) \Big\rangle,$$

(ii) *if $2m = n$, there is an equivalence $\mathcal{D}^b(coh(\mathcal{B})) \xrightarrow{\sim} \mathcal{D}^b(X)$.*

For $m = 0$, i.e. when there is no quadrics, the theorem coincides with Beilinson's description of the derived category of the projective space [Be]. For $m = 1$, this is Kapranov's description of the derived category of the quadric [Kap].

For odd n, the element d generates the center of $\mathcal{B}$ over $\mathcal{O}_{\mathbb{P}(U)}$. Hence the spectrum of the center of $\mathcal{B}$ is a ramified double cover Y over $\mathbb{P}(U)$. Also $\mathcal{B}$ yields a coherent sheaf of algebras $\mathcal{B}'$ over Y, such that $coh(\mathcal{B}')$ is equivalent to $coh(\mathcal{B})$. For the above case of two even dimensional quadrics, $\mathcal{B}'$ is an Azumaya algebra over $Y = C$. Since Brauer group of Y (taken over an algebraically closed field of characteristic zero) is trivial, the category $coh(\mathcal{B}')$ is equivalent to $coh(\mathcal{O}_Y)$. Hence (6.1) follows from the theorem.

Furthermore, when X is a K3 surface, the smooth intersection of 3 quadrics in $\mathbb{P}^5$, then the double cover Y is also a K3 surface, but $\mathcal{B}'$ is in general a nontrivial Azumaya algebra over Y. The theorem states an equivalence $\mathcal{D}^b(X) \simeq \mathcal{D}^b(coh(\mathcal{B}'))$.

This theorem illustrates the principle from section 3. The fully faithful functor is related to the moduli space of vector bundles on X, which are the restrictions to X of the spinor bundles on the quadrics. The (commutative) moduli space involved is either $\mathbb{P}(U)$ or Y, depending on parity of n.

Algebraically, the fully faithful functor in the theorem is given by an appropriate version of Koszul duality. Theorem 12 has a generalization to a class of Koszul Gorenstein algebras, which includes the coordinate rings of superprojective spaces.

References

[AKMW] Abramovich D., Karu K., Matsuki K., Wlodarczyk J., Torification and Factorization of Birational Maps *preprint* math.AG/9904135.

[Be] Beilinson A., Coherent sheaves on $\mathbb{P}^n$ and problems of linear algebra, *Funkcionalnyi analiz i ego pril.* 12 (1978), 68–69.

[Bo] Bondal A., Representations of associative algebras and coherent sheaves, *Izv. Akad. Nauk SSSR, Ser.Mat.* 53 (1989), 25–44; English transl. in *Math. USSR Izv.* 34 (1990).

[BK] Bondal A., Kapranov M., Representable functors, Serre functors, and mutations, *Izv. Akad. Nauk SSSR, Ser.Mat.*, 53 (1989), 1183–1205; English transl. in *Math. USSR Izv.*, 35 (1990), 519–541.

[BO1] Bondal A., Orlov D., Semiorthogonal decomposition for algebraic varieties, *preprint MPIM* 95/15 (1995), *preprint* math.AG/9506012.

[BO2] Bondal A., Orlov D., Reconstruction of a variety from the derived category and groups of autoequivalences, *Compositio Mathematica*, v.125 (2001) N.3, 327–344.

[BVdB] Bondal A., Van den Bergh M., Generators and representability of functors in commutative and noncommutative geometry, *preprint* math.AG/0204218.

[Br] Bridgeland T., Flops and derived categories, *preprint* math.AG/0009053.

[BKR] Bridgeland T., King A., Reid M., Mukai implies McKay: the McKay correspondence as an equivalence of derived categories, *J. Amer. Math. Soc.*, 14 (2001), 535–554, *preprint* math.AG/9908027.

[Kap] Kapranov M., On the derived categories of coherent sheaves on some homogeneous spaces, *Invent. Math.*, 92 (1988), 479–508.

[Kaw] Kawamata Y., Francia's flip and derived categories, *preprint* math.AG/0111041.

[Mu1] Mukai S., Duality between $D(X)$ and $D(\hat{X})$ with its application to Picard sheaves, *Nagoya Math. J.* 81 (1981), 153–175.

[Mu2] Mukai S., On the moduli space of bundles on a K3 surface I, *Vector bundles on algebraic varieties*, Tata Institute of Fundamental Research, Oxford University Press, Bombay and London, 1987.

[Na] Namikawa Y., Mukai flops and derived categories, *preprint* math.AG/0203287.

[Or1] Orlov D., Projective bundles, monoidal transformations and derived categories of coherent sheaves, *Izv. Akad. Nauk SSSR Ser.Mat.* 56 (1992), 852-862; English transl. in *Math. USSR Izv.* 38 (1993), 133–141.

[Or2] Orlov D., Equivalences of derived categories and K3 surfaces, *J. of Math. Sciences, Alg. geom.-5*, 84, N5, (1997), 1361–1381, *preprint* math.AG/9606006.

[Or3] Orlov D., On derived categories of coherent sheaves on abelian varieties and equivalences between them, *Izv. RAN, Ser. Mat.*, 66 (2002) N3, 131–158, (see also math.AG/9712017).

[Po] Polishchuk A., Symplectic biextensions and a generalization of the Fourier-Mukai transform, *Math. Res. Let.*, v.3 (1996), 813–828.

[Ta] Tate J., Homology of Noetherian rings and local rings, *Illinois J. Math.*, 1 (1957) N1, 14–27.

[VdB] Van den Bergh M., Three-dimensional flops and non-commutative ring, *paper in preparation.*

[Wl] Wlodarczyk J., Birational cobordisms and factorization of birational maps, *preprint* math.AG/9904074.

ICM 2002 · Vol. II · 57–66

Algebraic Cobordism

M. Levine*

Abstract

Together with F. Morel, we have constructed in [6, 7, 8] a theory of *algebraic cobordism*, an algebro-geometric version of the topological theory of complex cobordism. In this paper, we give a survey of the construction and main results of this theory; in the final section, we propose a candidate for a theory of higher algebraic cobordism, which hopefully agrees with the cohomology theory represented by the $\mathbb{P}^1$-spectrum MGL in the Morel-Voevodsky stable homotopy category.

2000 Mathematics Subject Classification: 19E15, 14C99, 14C25.
Keywords and Phrases: Cobordism, Chow ring , K-theory.

1. Oriented cohomology theories

Fix a field k and let $\mathbf{Sch}_k$ denote the category of separated finite-type k-schemes. We let $\mathbf{Sm}_k$ be the full subcategory of smooth quasi-projective k-schemes.

We have described in [7] the notion of an *oriented cohomology theory* on $\mathbf{Sm}_k$. Roughly speaking, such a theory A^* consists of a contravariant functor from $\mathbf{Sm}_k$ to graded rings (commutative), which is also covariantly functorial for projective equi-dimensional morphisms $f : Y \to X$ (with a shift in the grading):

$$f_* : \mathbb{A}^*(Y) \to A^{*-\dim_X Y}(X).$$

The pull-back g^* and push-forward f_* satisfy a projection formula and commute in transverse cartesian squares. If $L \to X$ is a line bundle with zero-section $s : X \to L$, we have the *first Chern class* of L, defined by

$$c_1(L) := s^*(s_*(1_X)) \in A^1(X),$$

where $1_X \in A^0(X)$ is the unit. A^* satisfies the *projective bundle formula*:

*Department of Mathematics, Northeastern University, Boston, MA 02115, USA. E-mail: marc@neu.edu

(PB) Let $\mathcal{E}$ be a rank $r+1$ locally free coherent sheaf on X, with projective bundle $q : \mathbb{P}(\mathcal{E}) \to X$ and tautological quotient invertible sheaf $q^*\mathcal{E} \to \mathcal{O}(1)$. Let $\xi = c_1(\mathcal{O}(1))$. Then $A^*(\mathbb{P}(\mathcal{E}))$ is a free $A^*(X)$-module with basis $1, \xi, \ldots, \xi^r$.

Finally, A^* satisfies a homotopy property: if $p : V \to X$ is an affine-space bundle (i.e., a torsor for a vector bundle over X), then $p^* : A^*(X) \to A^*(V)$ is an isomorphism.

Examples 1.1. (1) The theories CH^* and $H^{2*}_{\text{ét}}(-, \mathbb{Z}/n(*))$ on $\mathbf{Sm}_k$ (also with $\mathbb{Z}_l(*)$ or $\mathbb{Q}_l(*)$ coefficients).

(2) The theory $K_0[\beta, \beta^{-1}]$ on $\mathbf{Sm}_k$. Here β is an indeterminant of degree -1, used to keep track of the relative dimension when taking projective push-forward.

Remarks 1.2. (1) In [8], we consider a more general (dual) notion, that of an *oriented Borel-Moore homology theory* A_*. Roughly, this is a functor from a full subcategory of $\mathbf{Sch}_k$ to graded abelian groups, covariant for projective maps, and contravariant (with a shift in the grading) for local complete intersection morphisms. In addition, one has external products, and a degree -1 Chern class endomorphism $\tilde{c}_1(L) : A_*(X) \to A_{*-1}(X)$ for each line bundle L on X, defined by $\tilde{c}_1(L)(\eta) = s^*(s_*(\eta))$, $s : X \to L$ the zero-section. As for an oriented cohomology theory, there are various compatibilities of push-forward and pull-back, and A_* satisfies a projective bundle formula and a homotopy property.

This allows for a more general category of definition for A_*, e.g., the category $\mathbf{Sch}_k$. As we shall see, the setting of Borel-Moore homology is often more natural than cohomology. On $\mathbf{Sm}_k$, the two notions are equivalent: to pass from Borel-Moore homology to cohomology, one re-grades by setting $A^n(X) := A_{n-\dim_k X}(X)$ and uses the l.c.i. pull-back for A_* to give the contravariant functoriality of A^*, noting that every morphism of smooth k-schemes is an l.c.i. morphism. We will state most of our results for cohomology theories on $\mathbf{Sm}_k$, but they extend to the setting of Borel-Moore homology on $\mathbf{Sch}_k$ (see [8] for details).

(2) Our notion of oriented cohomology is related to that of Panin [10], but is not the same.

2. The formal group law

Let A_* be an oriented cohomology theory on $\mathbf{Sm}_k$. As noticed by Quillen [11], a double application of the projective bundle formula (PB) yields the isomorphism of rings

$$A^*(k)[[u, v]] \cong \varprojlim_{n,m} A^*(\mathbb{P}^n \times \mathbb{P}^m),$$

the isomorphism sending u to $c_1(p_1^*\mathcal{O}(1))$ and v to $c_1(p_2^*\mathcal{O}(1))$. The class of $c_1(p_1^*\mathcal{O}(1) \otimes p_2^*\mathcal{O}(1))$ thus gives a power series $F_A(u, v) \in A^*(k)[[u, v]]$ with

$$c_1(p_1^*\mathcal{O}(1) \otimes p_2^*\mathcal{O}(1)) = F_A(c_1(p_1^*\mathcal{O}(1)), c_1(p_2^*\mathcal{O}(1))).$$

By the naturality of c_1, we have the identity for $X \in \mathbf{Sm}_k$ with line bundles L, M,

$$c_1(L \otimes M) = F_A(c_1(L), c_1(M)).$$

In addition, $F_A(u,v) = u+v \mod uv$, $F_A(u,v) = F_A(v,u)$, and $F_A(F_A(u,v),w) = F_A(u,F_A(v,w))$. Thus, F_A gives a formal group law with coefficients in $A^*(k)$.

Remark 2.3. Note that $c_1 : \mathrm{Pic}(X) \to A^1(X)$ is a group homomorphism if and only if $F_A(u,v) = u+v$. If this is the case, we call A^* *ordinary*, if not, A^* is *extraordinary*. If $F_A(u,v) = u+v-\alpha uv$ with α a unit in $A^*(k)$, we call A^* *multiplicative and periodic.*

Examples 2.4. For $A^* = \mathrm{CH}^*$ or H^{2*}, $F_A = u+v$, giving examples of ordinary theories. For the theory $A = K_0[\beta,\beta^{-1}]$, $c_1(L) = (1-L^\vee)\beta^{-1}$, and $F_A(u,v) = u+v-\beta uv$, giving an example of a multiplicative and periodic theory.

Remark 2.5. Let $\tilde{\mathbb{L}}^* = \mathbb{Z}[a_{ij} \mid i,j \geq 1]$, where we give a_{ij} degree $-i-j+1$, and let $F \in \tilde{\mathbb{L}}^*[[u,v]]$ be the power series $F = u+v+\sum_{ij} a_{ij}u^iv^j$. Let

$$\mathbb{L}^* = \tilde{\mathbb{L}}^*/F(u,v) = F(v,u), F(F(u,v),w) = F(u,F(v,w)),$$

and let $F_{\mathbb{L}} \in \mathbb{L}^*[[u,v]]$ be the image of F. Then $(F_{\mathbb{L}}, \mathbb{L}^*)$ is the universal commutative dimension 1 formal group; $\mathbb{L}^*$ is called the *Lazard ring* (*cf.* [5]).

Thus, if A^* is an oriented cohomology theory on $\mathbf{Sm}_k$, there is a canonical graded ring homomorphism $\phi_A : \mathbb{L}^* \to A^*(k)$ with $\phi_A(F_{\mathbb{L}}) = F_A$.

3. Algebraic cobordism

The main result of [7, 8] is

Theorem 3.6. *Let k be a field of characteristic zero.*

1. *There is a universal oriented Borel-Moore homology theory Ω_* on $\mathbf{Sch}_k$. The restriction of Ω_* to $\mathbf{Sm}_k$ yields the universal oriented cohomology theory Ω^* on $\mathbf{Sm}_k$.*
2. *The homomorphism $\phi_\Omega : \mathbb{L}^* \to \Omega^*(k)$ is an isomorphism.*
3. *Let $i : Z \to X$ be a closed imbedding with open complement $j : U \to X$. Then the sequence*

$$\Omega_*(Z) \xrightarrow{i_*} \Omega_*(X) \xrightarrow{j^*} \Omega_*(U) \to 0$$

is exact.

Idea of construction: We construct $\Omega_*(X)$ in steps; the construction is inspired by Quillen's approach to complex cobordism [11].

1. Start with *cobordism cycles* $(f : Y \to X, L_1, \ldots, L_r)$, with $Y \in \mathbf{Sm}_k$ irreducible, $f : Y \to X$ projective and $L_1, \ldots, L_r$ line bundles on Y (we allow $r = 0$). We identify two cobordism cycles if there is an isomorphism $\phi : Y \to Y'$, a permutation σ and isomorphisms $L_j \cong \phi^* L'_{\sigma(j)}$. Let $\mathcal{Z}_*(X)$ be the free abelian group on the cobordism cycles, graded by giving $(f : Y \to X, L_1, \ldots, L_r)$ degree $\dim_k Y - r$.

2. Let $\mathcal{R}^{dim}(X)$ be the subgroup of $\mathcal{Z}_*(X)$ generated by cobordism cycles of the form $(f : Y \to X, \pi^*L_1, \ldots, \pi^*L_r, M_1, \ldots, M_s)$, where $\pi : Y \to Z$ is a smooth morphism in $\mathbf{Sm}_k$, the L_i are line bundles on Z, and $r > \dim_k Z$. Let $\underline{\mathcal{Z}}_*(X) = \mathcal{Z}_*(X)/\mathcal{R}^{dim}(X)$.
3. Add the Gysin isomorphism: If $L \to Y$ is a line bundle and $s : Y \to L$ is a section transverse to the zero-section with divisor $i : D \to Y$, identify $(f : Y \to X, L_1, \ldots, L_r, L)$ with $(f \circ i : D \to X, i^*L_1, \ldots, i^*L_r)$. We let $\underline{\Omega}_*(X)$ denote the resulting quotient of $\underline{\mathcal{Z}}_*(X)$. Note that on $\underline{\Omega}_*(X)$ we have, for each line bundle $L \to X$, the *Chern class operator*

$$\tilde{c}_1(L) : \underline{\Omega}_*(X) \to \underline{\Omega}_{*-1}(X)$$
$$(f : Y \to X, L_1, \ldots, L_r) \mapsto (f : Y \to X, L_1, \ldots, L_r, f^*L)$$

as well as push-forward maps $f_* : \underline{\Omega}_*(X) \to \underline{\Omega}_*(X')$ for $f : X \to X'$ projective.
4. Impose the formal group law: Regrade $\mathbb{L}$ by setting $\mathbb{L}_n := \mathbb{L}^{-n}$. Let $\Omega_*(X)$ be the quotient of $\mathbb{L}_* \otimes \underline{\Omega}_*(X)$ by the imposing the identity of maps $\mathbb{L}_* \otimes \underline{\Omega}_*(Y) \to \mathbb{L}_* \otimes \underline{\Omega}_*(X)$

$$(\mathrm{id} \otimes f_*) \circ F_{\mathbb{L}}(\tilde{c}_1(L), \tilde{c}_1(M)) = \mathrm{id} \otimes (f_* \circ \tilde{c}_1(L \otimes M))$$

for $f : Y \to X$ projective, and L, M line bundles on Y. Note that, having imposed the relations in $\mathcal{R}^{dim}$, the operators $\tilde{c}_1(L)$, $\tilde{c}_1(M)$ are locally nilpotent, so the infinite series $F_{\mathbb{L}}(\tilde{c}_1(L), \tilde{c}_1(M))$ makes sense.

As the notation suggests, the most natural construction of Ω is as an oriented Borel-Moore homology theory rather than an oriented cohomology theory; the tranlation to an oriented cohomology theory on $\mathbf{Sm}_k$ is given as in remark 1.2(1). The proof of theorem 3.6 uses resolution of singularities [4] and the weak factorization theorem [1] in an essential way.

Remark 3.7. In addition to the properties of Ω_* listed in theorem 3.6, $\Omega_*(X)$ is generated by the classes of "elementary" cobordism cycles $(f : Y \to X)$.

4. Degree formulas

In the paper [12], Rost made a number of conjectures based on the theory of algebraic cobordism in the Morel-Voevodsky stable homotopy category. Many of Rost's conjectures have been proved by homotopy-theoretic means (see [3]); our construction of algebraic cobordism gives an alternate proof of these results, and settles many of the remaining open questions as well. We give a sampling of some of these results.

4.1. The generalized degree formula

All the degree formulas follow from the "generalized degree formula". We first define the degree map $\Omega^*(X) \to \Omega^*(k)$.

Definition 4.8. Let k be a field of characteristic zero and let X be an irreducible finite type k-scheme with generic point $i : x \to X$. For an element η of $\Omega^*(X)$, define $\deg \eta \in \Omega^*(k)$ to be the element mapping to $i^*\eta$ in $\Omega^*(k(x))$ under the isomorphisms $\Omega^*(k) \cong \mathbb{L}^* \cong \Omega^*(k(x))$ given by theorem 3.6(2).

Theorem 4.9(generalized degree formula). *Let k be a field of characteristic zero. Let X be an irreducible finite type k-scheme, and let η be in $\Omega_*(X)$. Let $f_0 : B_0 \to X$ be a resolution of singularities of X, with B_0 quasi-projective over k. Then there are $a_i \in \Omega_*(k)$, and projective morphisms $f_i : B_i \to X$ such that*

1. *Each B_i is in $\mathbf{Sm}_k$, $f_i : B_i \to f(B_i)$ is birational and $f(B_i)$ is a proper closed subset of X (for $i > 0$).*
2. *$\eta - (\deg \eta)[f_0 : B_0 \to X] = \sum_{i=1}^{r} a_i[f_i : B_i \to X]$ in $\Omega_*(X)$.*

Proof. It follows from the definitions of Ω^* that we have

$$\Omega^*(k(x)) = \varinjlim_{U} \Omega^*(U),$$

where the limit is over smooth dense open subschemes U of X, and $\Omega^*(k(x))$ is the value at $\operatorname{Spec} k(x)$ of the functor Ω^* on finite type $k(x)$-schemes. Thus, there is a smooth open subscheme $j : U \to X$ of X such that $j^*\eta = (\deg \eta)[\mathrm{id}_U]$ in $\Omega^*(U)$. Since $U \times_X B_0 \cong U$, it follows that $j^*(\eta - (\deg \eta)[f_0]) = 0$ in $\Omega^*(U)$.

Let $W = X \setminus U$. From the localization sequence

$$\Omega_*(W) \xrightarrow{i_*} \Omega_*(X) \xrightarrow{j^*} \Omega_*(U) \to 0,$$

we find an element $\eta_1 \in \Omega_*(W)$ with $i_*(\eta_1) = \eta - (\deg \eta)[f_0]$, and noetherian induction completes the proof. □

Remark 4.10. Applying theorem 4.9 to the class of a projective morphism $f : Y \to X$, with $X, Y \in \mathbf{Sm}_k$, we have the formula

$$[f : Y \to X] - (\deg f)[\mathrm{id}_X] = \sum_{i=1}^{r} a_i[f_i : B_i \to X]$$

in $\Omega^*(X)$. Also, if $\dim_k X = \dim_k Y$, $\deg f$ is the usual degree, i.e., the field extension degree $[k(Y) : k(X)]$ if f is dominant, or zero if f is not.

4.2. Complex cobordism

For a differentiable manifold M, one has the complex cobordism ring $MU^*(M)$. Given an embedding $\sigma : k \to \mathbb{C}$ and an $X \in \mathbf{Sm}_k$, we let $X^\sigma(\mathbb{C})$ denote the complex manifold associated to the smooth $\mathbb{C}$-scheme $X \times_k \mathbb{C}$. Sending X to $MU^{2*}(X^\sigma(\mathbb{C}))$ defines an oriented cohomology theory on $\mathbf{Sm}_k$; by the universality of Ω^*, we have a natural homomorphism

$$\Re_\sigma : \Omega^*(X) \to MU^{2*}(X^\sigma(\mathbb{C})).$$

Now, if $P = P(c_1, \ldots, c_d)$ is a degree d (weighted) homogeneous polynomial, it is known that the operation of sending a smooth compact d-dimensional complex manifold M to the Chern number $\deg(P(c_1, \ldots, c_d)(\Theta_M))$ (where Θ_M is the complex tangent bundle) descends to a homomorphism $MU^{-2d} \to \mathbb{Z}$. Composing with $\Re_\sigma$, we have the homomorphism $P : \Omega^{-d}(k) \to \mathbb{Z}$. If X is smooth and projective of dimension d over k, we have $P([X]) = \deg(P(c_1, \ldots, c_d)(\Theta_{X^\sigma(\mathbb{C})}))$; $P([X])$ is in fact independent of the choice of embedding σ.

Let $s_d(c_1, \ldots, c_d)$ be the polynomial which corresponds to $\sum_i \xi_i^d$, where $\xi_1, \ldots$ are the Chern roots. The following divisibility is known (see [2]): if $d = p^n - 1$ for some prime p, and $\dim X = d$, then $s_d(X)$ is divisible by p.

In addition, for integers $d = p^n - 1$ and $r \geq 1$, there are mod p characteristic classes $t_{d,r}$, with $t_{d,1} = s_d/p \mod p$. The s_d and the $t_{d,r}$ have the following properties:

(4.1)

1. $s_d(X) \in p\mathbb{Z}$ is defined for X smooth and projective of dimension $d = p^n - 1$. $t_{d,r}(X) \in \mathbb{Z}/p$ is defined for X smooth and projective of dimension $rd = r(p^n - 1)$.
2. s_d and $t_{d,r}$ extend to homomorphisms $s_d : \Omega^{-d}(k) \to p\mathbb{Z}$, $t_{d,r} : \Omega^{-rd}(k) \to \mathbb{Z}/p$.
3. If X and Y are smooth projective varieties with $\dim X, \dim Y > 0$, $\dim X + \dim Y = d$, then $s_d(X \times Y) = 0$.
4. If $X_1, \ldots, X_s$ are smooth projective varieties with $\sum_i \dim X_i = rd$, then $t_{d,r}(\prod_i X_i) = 0$ unless $d | \dim X_i$ for each i.

We can now state Rost's degree formula and the higher degree formula:

Theorem 4.11(Rost's degree formula). *Let $f : Y \to X$ be a morphism of smooth projective k-schemes of dimension d, $d = p^n - 1$ for some prime p. Then there is a zero-cycle η on X such that*

$$s_d(Y) - (\deg f) s_d(X) = p \cdot \deg(\eta).$$

Theorem 4.12(Rost's higher degree formula). *Let $f : Y \to X$ be a morphism of smooth projective k-schemes of dimension rd, $d = p^n - 1$ for some prime p. Suppose that X admits a sequence of surjective morphisms*

$$X = X_0 \to X_1 \to \ldots \to X_{r-1} \to X_r = \operatorname{Spec} k,$$

such that:

1. $\dim X_i = d(r - i)$.
2. *Let η be a zero-cycle on $X_i \times_{X_{i+1}} \operatorname{Spec} k(X_{i+1})$. Then $p | \deg(\eta)$.*

Then

$$t_{d,r}(Y) = \deg(f) t_{d,r}(X).$$

Proof. These two theorems follow easily from the generalized degree formula. Indeed, for theorem 4.11, take the identity of remark 4.10 and push forward to

$\Omega^*(k)$. Using remark 3.7, this gives the identity

$$[Y] - (\deg f)[X] = \sum_{i=1}^{r} m_i[A_i \times B_i]$$

in $\Omega^*(X)$, for smooth, projective k-schemes A_j, B_j, and integers m_j, where each B_i admits a projective morphism $f_i : B_i \to X$ which is birational to its image and not dominant. Since s_d vanishes on non-trivial products, the only relevant part of the sum involves those B_j of dimension zero; such a B_j is identified with the closed point $b_j := f_j(B_j)$ of X. Applying s_d, we have

$$s_d(Y) - \deg(f)s_d(X) = \sum_j m_j s_d(A_j) \deg_k(b_j).$$

Since $s_d(A_j) = pn_j$ for suitable integers n_j, we have

$$s_d(Y) - \deg(f)s_d(X) = p \deg(\sum_j m_j n_j b_j).$$

Taking $\eta = \sum_j m_j n_j b_j$ proves theorem 4.11.

The proof of theorem 4.12 is similar: Start with the decomposition of $[f : Y \to X] - (\deg f)[\mathrm{id}_X]$ given by remark 4.10. One then decomposes the maps $B_i \to X = X_0$ further by pushing forward to X_1 and using theorem 4.9. Iterating down the tower gives the identity in $\Omega_*(k)$

$$[Y] - (\deg f)[X] = \sum_i m_i[B_0^i \times \ldots \times B_r^i];$$

the condition (2) implies that, if $d|\dim_k B_j^i$ for all $j = 0, \ldots, r$, then $p|m_j$. Applying $t_{d,r}$ and using the property (4.1)(4) yields the formula. □

5. Comparison results

Suppose we have a formal group (f, R), giving the canonical homomorphism $\phi_f : \mathbb{L}^* \to R$. Let $\Omega^*_{(f,R)}$ be the functor

$$\Omega^*_{(f,R)}(X) = \Omega^*(X) \otimes_{\mathbb{L}^*} R,$$

where $\Omega^*(X)$ is an $\mathbb{L}^*$-algebra via the homomorphism $\phi_\Omega : \mathbb{L}^* \to \Omega^*(k)$. The universal property of Ω^* gives the analogous universal property for $\Omega^*_{(f,R)}$.

In particular, let Ω^*_+ be the theory with $(f(u,v), R) = (u+v, \mathbb{Z})$, and let $\Omega^*_\times$ be the theory with $(f(u,v), R) = (u+v-\beta uv, \mathbb{Z}[\beta, \beta^{-1}])$. We thus have the canonical natural transformations of oriented theories on $\mathbf{Sm}_k$

$$\Omega^*_+ \to \mathrm{CH}^*; \quad \Omega^*_\times \to K_0[\beta, \beta^{-1}]. \tag{5.2}$$

Theorem 5.13. *Let k be a field of characteristic zero. The natural transformations* (5.2) *are isomorphisms, i.e.,* CH^* *is the universal ordinary oriented cohomology theory and* $K_0[\beta, \beta^{-1}]$ *is the universal multiplicative and periodic theory.*

Proof. For CH^*, this uses localization, theorem 4.9 and resolution of singularities. For K_0, one writes down an integral Chern character, which gives the inverse isomorphism by the Grothendieck-Riemann-Roch theorem. □

6. Higher algebraic cobordism

The cohomology theory represented by the $\mathbb{P}^1$-spectrum MGL in the Morel-Voevodsky $\mathbb{A}^1$-stable homotopy category [9, 13] gives perhaps the most natural algebraic analogue of complex cobordism. By universality, $\Omega^n(X)$ maps to $MGL^{2n,n}(X)$; to show that this map is an isomorphism, one would like to give a map in the other direction. For this, the most direct method would be to extend Ω^* to a theory of higher algebraic cobordism; we give one possible approach to this construction here.

The idea is to repeat the construction of Ω_*, replacing abelian groups with symmetric monoidal categories throughout. Comparing with the Q-construction, one sees that the cobordism cycles in $\mathcal{R}^{dim}(X)$ should be homotopic to zero, but not canonically so. Thus, we cannot impose this relation directly, forcing us to modify the group law by taking a limit.

Start with the category $\widetilde{\mathcal{Z}}(X)_0$, with objects $(f : Y \to X, L_1, \ldots, L_r)$, where Y is irreducible in $\mathbf{Sm}_k$, f is projective, and the L_i are line bundles on Y. A morphism $(f : Y \to X, L_1, \ldots, L_r) \to (f' : Y' \to X, L'_1, \ldots, L'_r)$ in $\widetilde{\mathcal{Z}}(X)_0$ consist of a tuple $(\phi, \psi_1, \ldots, \psi_r, \sigma)$, with $\phi : Y \to Y'$ an isomorphism over X, σ a permutation, and $\psi_j : L_j \to \phi^* L'_{\sigma(j)}$ an isomorphism of line bundles on Y. Form the category $\widetilde{\mathcal{Z}}(X)$ as the symmetric monoidal category freely generated by $\widetilde{\mathcal{Z}}(X)_0$; grade $\widetilde{\mathcal{Z}}(X)$ by letting $\widetilde{\mathcal{Z}}_n(X)$ be the full symmetric monoidal subcategory generated by the $(f : Y \to X, L_1, \ldots, L_r)$ with $n = \dim_k Y - r$.

Next, form $\underline{\widetilde{\Omega}}(X)$ by adjoining (as a symmetric monoidal category) an isomorphism $\gamma_{L,s} : (f \circ i : D \to X, i^*L_1, \ldots, i^*L_r) \to (f : Y \to X, L_1, \ldots, L_r, L)$ for each section $s : Y \to L$ transverse to the zero-section with divisor $i : D \to X$. Given a morphism $\tilde{\phi} := (\phi, \ldots) : (f : Y \to X, L_1, \ldots, L_r, L) \to (f' : Y' \to X, L'_1, \ldots, L'_r, L')$ (with $L \cong \phi^* L'$ via $\tilde{\phi}$), let $i' : D' \to Y'$ be the map induced by ϕ, $s' : Y' \to L'$ the section induced by s, and

$$\psi^D : (f \circ i : D \to X, i^*L_1, \ldots, i^*L_r) \to (f' \circ i' : D' \to X, i'^*L'_1, \ldots, i'^*L'_r),$$

the morphism induced by ψ. We impose the relation $\psi \circ \gamma_{L,s} = \gamma_{L',s'} \circ \psi^D$. Finally, for line bundles L, M with smooth transverse divisors $i_D : D \to Y$, $i_E : E \to Y$ defined by sections $s : Y \to L$, $t : Y \to M$, respectively, we impose the relation $\gamma_{L,s} \circ \gamma_{i_D^* M, i_D^* t} = \gamma_{M,t} \circ \gamma_{i_E^* L, i_E^* s}$. The grading on $\widetilde{\mathcal{Z}}(X)$ extends to one on $\underline{\widetilde{\Omega}}(X)$.

Given $g : X \to X'$ projective, we have the functor $g_* : \underline{\widetilde{\Omega}}(X) \to \underline{\widetilde{\Omega}}(X')$, similarly, given a smooth morphism $h : X \to X'$, we have the functor $h^* : \underline{\widetilde{\Omega}}(X') \to \underline{\widetilde{\Omega}}(X)$. Given a line bundle L on X, we have the natural transformation $\tilde{c}_1(L)$ sending $(f : Y \to X, L_1, \ldots, L_r)$ to $(f : Y \to X, L_1, \ldots, L_r, f^*L)$.

Now let $\mathcal{C}$ be a symmetric monoidal category such that all morphisms are isomorphisms, and let R be a ring, free as a $\mathbb{Z}$-module. One can define a symmetric monoidal category $R \otimes_{\mathbb{N}} \mathcal{C}$ with a symmetric monoical functor $\mathcal{C} \to R \otimes_{\mathbb{N}} \mathcal{C}$ which

is universal for symmetric monoidal functors $\mathcal{C} \to \mathcal{C}'$ such that $\mathcal{C}'$ admits an action of R via natural transformations. In case $R = \mathbb{Z}$, $\mathbb{Z} \otimes_{\mathbb{N}} \mathcal{C}$ is the standard group completion $\mathcal{C}^{-1}\mathcal{C}$. In general, if $\{e_\alpha \mid \alpha \in A\}$ is a $\mathbb{Z}$-basis for R, then

$$R \otimes_{\mathbb{N}} \mathcal{C} = \coprod_{\alpha} \mathcal{C}^{-1}\mathcal{C},$$

with the R-action given by expressing $\times x : R \to R$ in terms of the basis $\{e_\alpha\}$.

For each integer $n \geq 0$, let $\mathbb{L}_*^{(n)}$ be the quotient of $\mathbb{L}_*$ by the ideal of elements of degree $> n$. We thus have the formal group $(F_{\mathbb{L}^{(n)}}, \mathbb{L}_*^{(n)})$.

We form the category $\mathbb{L}^{(n)} \otimes_{\mathbb{N}} \widetilde{\underline{\Omega}}(X)$, which we grade by total degree. For each $f : Y \to X$ projective, with $Y \in \mathbf{Sm}_k$, and line bundles L, M, $L_1, \ldots, L_r$ on Y, we adjoin an isomophism $\rho_{L,M}$

$$f_*(F_{\mathbb{L}^{(n)}}(\tilde{c}_1(L), \tilde{c}_1(M))(\mathrm{id}_Y, L_1, \ldots, L_r)) \xrightarrow{\sim} f_*(\mathrm{id} \otimes \tilde{c}_1(L \otimes M)(\mathrm{id}_Y, L_1, \ldots, L_r)).$$

We impose the condition of naturality with respect to the maps in $\mathbb{L}^{(n)} \otimes_{\mathbb{N}} \widetilde{\underline{\Omega}}_n(Y)$, in the evident sense; the Chern class transformations extend in the obvious manner.

We impose the following commutativity condition: We have the evident isomorphism $t_{L,M} : F_{\mathbb{L}^{(n)}}(\tilde{c}_1(L), \tilde{c}_1(M)) \to F_{\mathbb{L}^{(n)}}(\tilde{c}_1(M), \tilde{c}_1(L))$ of natural transformations, as well as $\tau_{L,M} : \tilde{c}_1(L \otimes M) \to \tilde{c}_1(M \otimes L)$, the isomorphism induced by the symmetry $L \otimes M \cong M \otimes L$. Then we impose the identity $\tau_{L,M} \circ \rho_{L,M} = \rho_{M,L} \circ t_{L,M}$. We impose a similar identity between the associativity of the formal group law and the associativity of the tensor product of line bundles.

We also adjoin $a \cdot \tau_{L,M}$ for all $a \in \mathbb{L}^{(n)}$, with similar compatibilities as above, respecting the $\mathbb{L}^{(n)}$-action and sum. This forms the symmetric monoidal category $\widetilde{\Omega}^{(n)}(X)$, which inherits a grading from $\widetilde{\underline{\Omega}}(X)$. We have the inverse system of graded symmetric monoidal categories:

$$\ldots \to \widetilde{\Omega}^{(n+1)}(X) \to \widetilde{\Omega}^{(n)}(X) \to \ldots .$$

Definition 5.14. Set $\Omega^{(n)}_{m,r}(X) := \pi_r(B\widetilde{\Omega}^{(n)}_m(X))$ and $\Omega_{m,r}(X) := \varprojlim_n \Omega^{(n)}_{m,r}(X)$.

At present, we can only verify the following:

Theorem 5.15. *There is a natural isomorphism* $\Omega_{m,0}(X) \cong \Omega_m(X)$.

Proof. First note that $\pi_0(\widetilde{\mathcal{Z}}_m(X))$ is a commutative monoid with group completion $\mathcal{Z}_m(X)$. Next, the natural map $\pi_0(\widetilde{\underline{\Omega}}_*(X))^+ \to \underline{\Omega}_*(X)$ is surjective with kernel generated by the classes generating $\mathcal{R}^{dim}(X)$. Given such an element $\psi := (f : Y \to X, \pi^*L_1, \ldots, \pi^*L_r, M_1, \ldots, M_s)$, with $\pi : Y \to Z$ smooth, and $r > \dim_k Z$, suppose that the L_i are very ample. We may then choose sections $s_i : Z \to L_i$ with divisors D_i all intersecting transversely. Iterating the isomorphisms γ_{L_i,s_i} gives a path from ψ to 0 in $B\widetilde{\underline{\Omega}}_*(X)$. Passing to $B\widetilde{\Omega}^{(n)}_m(X)$, the group law allows us to replace an arbitrary line bundle witha difference of very ample ones, so all the classes of this form go to zero in $\Omega^{(n)}_{m,0}(X)$. This shows that the natural map

$$\Omega^{(n)}_{m,0}(X) \to (\mathbb{L}^{(n)} \otimes_{\mathbb{L}} \Omega_*(X))_m$$

is an isomorphism. Since $(\mathbb{L}^{(n)} \otimes_{\mathbb{L}} \Omega_*(X))_m = \Omega_m(X)$ for $m \geq n$, we are done. □

The categories $\widetilde{\Omega}_m^{(n)}(X)$ are covariantly functorial for projective maps, contravariant for smooth maps (with a shift in the grading) and have first Chern class natural transformations $\tilde{c}_1(L) : \widetilde{\Omega}_m^{(n)}(X) \to \widetilde{\Omega}_{m-1}^{(n)}(X)$ for $L \to X$ a line bundle.

We conjecture that the inverse system used to define $\Omega_{m,r}(X)$ is eventually constant for all r, not just for $r = 0$. If this is true, it is reasonable to define the space $B\widetilde{\Omega}_m(X)$ as the homotopy limit

$$B\widetilde{\Omega}_m(X) := \operatorname*{holim}_n B\widetilde{\Omega}_m^{(n)}(X).$$

One would then have $\Omega_{m,r}(X) = \pi_r(B\widetilde{\Omega}_m(X), 0)$ for all m, r; hopefully the properties of Ω_* listed in theorem 3.6 would then generalize into properties of the spaces $B\widetilde{\Omega}_m(X)$.

References

[1] D. Abramovich, K. Karu, K. Matsuki, J. Wlodarczyk, Torification and factorization of birational morphisms, preprint 2000, AG/9904135.

[2] J. F. Adams, *Stable homotopy and generalised homology*, Chicago Lectures in Mathematics. University of Chicago Press, Chicago, Ill.-London, 1974.

[3] S. Borghesi, Algebraic Morava K-theories and the higher degree formula, preprint May 2000, www.math.uiuc.edu/K-theory/0412/index.html.

[4] H. Hironaka, Resolution of singularities of an algebraic variety over a field of characteristic zero. I, II, *Ann. of Math.*, (2) 79 (1964), 109–203; ibid. 205–326.

[5] M. Lazard, Sur les groupes de Lie formels à un paramètre, *Bull. Soc. Math. France*, 83 (1955), 251-274.

[6] M. Levine et F. Morel, Cobordisme algébrique I, II, *C.R. Acad. Sci. Paris*, Série I, 332 (2001), 723-728; ibid. 815-820.

[7] M. Levine et F. Morel, Algebraic cobordism, I, preprint Feb. 2002, www.math.uiuc.edu/K-theory/0547/index.html.

[8] M. Levine, Algebraic cobordism, II, preprint June 2002, www.math.uiuc.edu/K-theory/0577/index.html.

[9] F. Morel, V. Voevodsky, $\mathbb{A}^1$ *homotopy of schemes*, Publications Mathématiques de l'I.H.E.S, volume 90.

[10] I. Panin, Push-forwards in oriented cohomology theories of algebraic varieties, preprint Nov. 2000, www.math.uiuc.edu/K-theory/0459/index.html.

[11] D. Quillen, Elementary proofs of some results of cobordism theory using Steenrod operations, *Advances in Math.*, 7 (1971), 29–56.

[12] M. Rost, Construction of splitting varieties, preprint, 1998.

[13] V. Voevodsky, $\mathbb{A}^1$-homotopy theory, *Proceedings of the International Congress of Mathematicians*, Vol. I (1998). *Doc. Math.* Extra Vol. I (1998), 579–604.

ICM 2002 · Vol. II · 67–76

Permutation Groups and Normal Subgroups

Cheryl E. Praeger*

Abstract

Various descending chains of subgroups of a finite permutation group can be used to define a sequence of 'basic' permutation groups that are analogues of composition factors for abstract finite groups. Primitive groups have been the traditional choice for this purpose, but some combinatorial applications require different kinds of basic groups, such as quasiprimitive groups, that are defined by properties of their normal subgroups. Quasiprimitive groups admit similar analyses to primitive groups, share many of their properties, and have been used successfully, for example to study s-arc transitive graphs. Moreover investigating them has led to new results about finite simple groups.

2000 Mathematics Subject Classification: 20B05, 20B10 20B25, 05C25.
Keywords and Phrases: Automorphism group, Simple group, Primitive permutation group, Quasiprimitive permutation group, Arc-transitive graph.

1. Introduction

For a satisfactory understanding of finite groups it is important to study simple groups and characteristically simple groups, and how to fit them together to form arbitrary finite groups. This paper discusses an analogous programme for studying finite permutation groups. By considering various descending subgroup chains of finite permutation groups we define in §2 sequences of 'basic' permutation groups that play the role for finite permutation groups that composition factors or chief factors play for abstract finite groups. Primitive groups have been the traditional choice for basic permutation groups, but for some combinatorial applications larger families of basic groups, such as quasiprimitive groups, are needed (see §3).

Application of a theorem first stated independently in 1979 by M. E. O'Nan and L. L. Scott [4] has proved to be the most useful modern method for identifying the possible structures of finite primitive groups, and is now used routinely for their

*Department of Mathematics & Statistics, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia. E-mail: praeger@maths.uwa.edu.au

analysis. Analogues of this theorem are available for the alternative families of basic permutation groups. These theorems have become standard tools for studying finite combinatorial structures such as vertex-transitive graphs and examples are given in §3 of successful analyses for distance transitive graphs and s-arc-transitive graphs. Some characteristic properties of basic permutation groups, including these structure theorems are discussed in §4.

Studying the symmetry of a family of finite algebraic or combinatorial systems often leads to problems about groups of automorphisms acting as basic permutation groups on points or vertices. In particular determining the full automorphism group of such a system sometimes requires a knowledge of the permutation groups containing a given basic permutation group, and for this it is important to understand the lattice of basic permutation groups on a given set. The fundamental problem here is that of classifying all inclusions of one basic permutation group in another, and integral to its solution is a proper understanding of the factorisations of simple and characteristically simple groups. In §3 and §4 we outline the current status of our knowledge about such inclusions and their use.

The precision of our current knowledge of basic permutation groups depends heavily on the classification of the finite simple groups. Some problems about basic permutation groups translate directly to questions about simple groups, and answering them leads to new results about simple groups. Several of these results and their connections with basic groups are discussed in the final section §5.

In summary, this approach to analysing finite permutation groups involves an interplay between combinatorics, group actions, and the theory of finite simple groups. One measure of its success is its effectiveness in combinatorial applications.

2. Defining basic permutation groups

Let G be a subgroup of the symmetric group $\mathrm{Sym}(\Omega)$ of all permutations of a finite set Ω. Since an intransitive permutation group is contained in the direct product of its transitive constituents, it is natural when studying permutation groups to focus first on the transitive ones. Thus we will assume that G is transitive on Ω. Choose a point $\alpha \in \Omega$ and let G_α denote the subgroup of G of permutations that fix α, that is, the stabiliser of α. Let $\mathsf{Sub}(G, G_\alpha)$ denote the lattice of subgroups of G containing G_α. The concepts introduced below are independent of the choice of α because of the transitivity of G. We shall introduce three types of basic permutation groups, relative to $\mathcal{L}_1 := \mathsf{Sub}(G, G_\alpha)$ and two other types of lattices $\mathcal{L}_2$ and $\mathcal{L}_3$, where we regard each $\mathcal{L}_i$ as a function that can be evaluated on any finite transitive group G and stabiliser G_α.

For $G_\alpha \leq H \leq G$, the H-orbit containing α is $\alpha^H = \{\alpha^h \mid h \in H\}$. If $G_\alpha \leq H < K \leq G$, then the K-images of α^H form the parts of a K-invariant partition $\mathcal{P}(K, H)$ of α^K, and K induces a transitive permutation group $\mathsf{Comp}(K, H)$ on $\mathcal{P}(K, H)$ called a *component* of G. In particular the component $\mathsf{Comp}(G, G_\alpha)$ permutes $\mathcal{P}(G, G_\alpha) = \{\{\beta\} \mid \beta \in \Omega\}$ in the same way that G permutes Ω, and we may identify G with $\mathsf{Comp}(G, G_\alpha)$.

For a lattice $\mathcal{L}$ of subgroups of G containing G_α, we say that K *covers* H

in $\mathcal{L}$ if $K, H \in \mathcal{L}$, $H < K$, and there are no intermediate subgroups lying in $\mathcal{L}$. The *basic components of* G *relative to* $\mathcal{L}$ are then defined as all the components $\mathsf{Comp}(K, H)$ for which K covers H in $\mathcal{L}$. Each maximal chain $G_\alpha = G_0 < G_1 < \cdots < G_r = G$ in $\mathcal{L}$ determines a sequence of basic components relative to $\mathcal{L}$, namely $\mathsf{Comp}(G_1, G_0), \ldots, \mathsf{Comp}(G_r, G_{r-1})$, and G can be embedded in the iterated wreath product of these groups. In this way the permutation groups occurring as basic components relative to $\mathcal{L}$, for some finite transitive group, may be considered as 'building blocks' for finite permutation groups. We refer to such groups as basic permutation groups relative to $\mathcal{L}$.

A transitive permutation group G on Ω is *primitive* if G_α is a maximal subgroup of G, that is, if $\mathsf{Sub}(G, G_\alpha) = \{G, G_\alpha\}$. The basic components of G relative to $\mathcal{L}_1 = \mathsf{Sub}(G, G_\alpha)$ are precisely those of its components that are primitive.

The basic groups of the second type are the quasiprimitive groups. A transitive permutation group G on Ω is *quasiprimitive* if each nontrivial normal subgroup of G is transitive on Ω. The corresponding sublattice is the set $\mathcal{L}_2$ of all subgroups $H \in \mathsf{Sub}(G, G_\alpha)$ such that there is a sequence $H_0 = H \leq H_1 \leq \cdots \leq H_r = G$ with each subgroup of the form $H_i = G_\alpha N_i$ where for $i < r$, N_i is a normal subgroup of H_{i+1}, and $N_r = G$. The basic components of G relative to $\mathcal{L}_2$ are precisely those of its components that are quasiprimitive.

Basic groups of the third type are *innately transitive*, namely transitive permutation groups that have at least one transitive minimal normal subgroup. The corresponding sublattice will be $\mathcal{L}_3$. A subgroup N of G is *subnormal* in G if there is a sequence $N_0 = N \leq N_1 \leq \cdots \leq N_r = G$ such that, for $i < r$, N_i is a normal subgroup of N_{i+1}. The lattice $\mathcal{L}_3$ consists of all subgroups of the form $G_\alpha N$, where N is subnormal in G and normalised by G_α. All the basic components of G relative to $\mathcal{L}_3$ are innately transitive. Note that each primitive group is quasiprimitive and each quasiprimitive group is innately transitive. Proofs of the assertions about $\mathcal{L}_2$ and $\mathcal{L}_3$ and their components may be found in [27].

3. The role of basic groups in graph theory

For many group theoretic and combinatorial applications finite primitive permutation groups are the appropriate basic permutation groups, since many problems concerning finite permutation groups can be reduced to the case of primitive groups. However such reductions are sometimes not possible when studying point-transitive automorphism groups of finite combinatorial structures because the components of the given point-transitive group have no interpretation as point-transitive automorphism groups of structures within the family under investigation. The principal motivation for studying some of these alternative basic groups came from graph theory, notably the study of s-arc transitive graphs ($s \geq 2$).

A finite graph $\Gamma = (\Omega, E)$ consists of a finite set Ω of points, called vertices, and a subset E of unordered pairs from Ω called edges. For $s \geq 1$, an *s-arc* of Γ is a vertex sequence $(\alpha_0, \alpha_1, \ldots, \alpha_s)$ such that each $\{\alpha_i, \alpha_{i+1}\}$ is an edge and $\alpha_{i-1} \neq \alpha_{i+1}$ for all i. We usually call a 1-arc simply an arc. Automorphisms of Γ are permutations of Ω that leave E invariant, and a subgroup G of the automorphism group $\mathrm{Aut}(\Gamma)$ is

s-arc-transitive if G is transitive on the s-arcs of Γ. If Γ is connected and is regular of *valency* $k > 0$ so that each vertex is in k edges, then an s-arc-transitive subgroup $G \leq \mathrm{Aut}(\Gamma)$ is in particular transitive on Ω and also, if $s \geq 2$, on $(s-1)$-arcs. It is natural to ask which of the components of this transitive permutation group G on Ω act as s-arc-transitive automorphism groups of graphs related to Γ.

For $G_\alpha \leq H \leq G$, there is a naturally defined quotient graph Γ_H with vertex set the partition of Ω formed by the G-images of the set α^H, where two such G-images are adjacent in Γ_H if at least one vertex in the first is adjacent to at least one vertex of the second. If Γ is connected and G is arc-transitive, then Γ_H is connected and G induces an arc-transitive automorphism group of Γ_H, namely the component $\mathsf{Comp}(G, H)$. If H is a maximal subgroup of G, then $\mathsf{Comp}(G, H)$ is both vertex-primitive and arc-transitive on Γ_H. This observation enables many questions about arc-transitive graphs to be reduced to the vertex-primitive case.

Perhaps the most striking example is provided by the family of finite distance transitive graphs. The *distance* between two vertices is the minimum number of edges in a path joining them, and G is *distance transive* on Γ if for each i, G is transitive on the set of ordered pairs of vertices at distance i. In particular if G is distance transitive on Γ then Γ is connected and regular, of valency k say. If $k = 2$ then Γ is a cycle and all cycles are distance transitive, so suppose that $k \geq 3$. If Γ_H has more than two vertices, then $\mathsf{Comp}(G, H)$ is distance transitive on Γ_H, while if Γ_H has only two vertices then H is distance transitive on a smaller graph Γ_2, namely Γ_2 has α^H as vertex set with two vertices adjacent if and only if they are at distance 2 in Γ (see for example [12]). Passing to Γ_H or Γ_2 respectively and repeating this process, we reduce to a vertex-primitive distance transitive graph. The programme of classifying the finite vertex-primitive distance transitive graphs is approaching completion, and surveys of progress up to the mid 1990's can be found in [12, 31]. The initial result that suggested a classification might be possible is the following. Here a group G is *almost simple* if $T \leq G \leq \mathrm{Aut}(T)$ for some nonabelian simple group T, and a permutation group G has *affine type* if G has an elementary abelian regular normal subgroup.

Theorem 3.1 [28] *If G is vertex-primitive and distance transitive on a finite graph Γ, then either Γ is known explicitly, or G is almost simple, or G has affine type.*

In general, if G is s-arc-transitive on Γ with $s \geq 2$, then none of the components $\mathsf{Comp}(G, H)$ with $G_\alpha < H < G$ is s-arc-transitive on Γ_H, so there is no hope that the problem of classifying finite s-arc-transitive graphs, or even giving a useful description of their structure, can be reduced to the case of vertex-primitive s-arc-transitive graphs. However the class of s-arc transitive graphs behaves nicely with respect to *normal quotients*, that is, quotients Γ_H where $H = G_\alpha N$ for some normal subgroup N of G. For such quotients, the vertex set of Γ_H is the set of N-orbits, G acts s-arc-transitively on Γ_H, and if Γ_H has more than two vertices then Γ is a cover of Γ_H in the sense that, for two N-orbits adjacent in Γ_H, each vertex in one N-orbit is adjacent in Γ to exactly one vertex in the other N-orbit. We say that Γ is a *normal cover* of Γ_H. If in addition N is a maximal intransitive normal subgroup of G with more than two orbits, then G is both vertex-quasiprimitive and s-arc-transitive on Γ_H, see [24]. If some quotient Γ_H has two vertices then Γ is bipartite,

and such graphs require a specialised analysis that parallels the one described here. On the other hand if Γ is not bipartite then Γ is a normal cover of at least one Γ_H on which the G-action is both vertex-quasiprimitive and s-arc-transitive. The wish to understand quasiprimitive s-arc transitive graphs led to the development of a theory for finite quasiprimitive permutation groups similar to the theory of finite primitive groups. Applying this theory led to a result similar to Theorem 3.1, featuring two additional types of quasiprimitive groups, called *twisted wreath type* and *product action type.* Descriptions of these types may be found in [24] and [25].

Theorem 3.2 [24] *If G is vertex-quasiprimitive and s-arc-transitive on a finite graph Γ with $s \geq 2$, then G is almost simple, or of affine, twisted wreath or product action type.*

Examples exist for each of the four quasiprimitive types, and moreover this division of vertex-quasiprimitive s-arc transitive graphs into four types has resulted in a better understanding of these graphs, and in some cases complete classifications. For example all examples with G of affine type, or with $T \leq G \leq \mathrm{Aut}(T)$ and $T = \mathrm{PSL}_2(q), \mathrm{Sz}(q)$ or $\mathrm{Ree}(q)$ have been classified, in each case yielding new s-arc transitive graphs, see [13, 25]. Also using Theorem 3.2 to study the normal quotients of an s-arc transitive graph has led to some interesting restrictions on the number of vertices.

Theorem 3.3 [15, 16] *Suppose that Γ is a finite s-arc-transitive graph with $s \geq 4$. Then the number of vertices is even and not a power of 2.*

The concept of a normal quotient has proved useful for analysing many families of edge-transitive graphs, even those for which a given edge-transitive group is not vertex-transitive. For example it provides a framework for a systematic study of locally s-arc-transitive graphs in which quasiprimitive actions are of central importance, see [11].

We have described how to form primitive arc-transitive quotients of arc-transitive graphs, and quasiprimitive s-arc-transitive normal quotients of non-bipartite s-arc-transitive graphs. However recognising these quotients is not always easy without knowing their full automorphism groups. To identify the automorphism group of a graph, given a primitive or quasiprimitive subgroup G of automorphisms, it is important to know the permutation groups of the vertex set that contain G, that is the over-groups of G. In the case of finite primitive arc-transitive and edge-transitive graphs, knowledge of the lattice of primitive permutation groups on the vertex set together with detailed knowledge of finite simple groups led to the following result. The *socle* of a finite group G, denoted $\mathrm{soc}(G)$, is the product of its minimal normal subgroups.

Theorem 3.4 [22] *Let G be a primitive arc- or edge-transitive group of automorphisms of a finite connected graph Γ. Then either G and $\mathrm{Aut}(\Gamma)$ have the same socle, or $G < H \leq \mathrm{Aut}(\Gamma)$ where $\mathrm{soc}(G) \neq \mathrm{soc}(H)$ and G, H are explicitly listed.*

In the case of graphs Γ for which a quasiprimitive subgroup G of $\mathrm{Aut}(\Gamma)$ is given, it is possible that $\mathrm{Aut}(\Gamma)$ may not be quasiprimitive. However, even in this

case a good knowledge of the quasiprimitive over-groups of a quasiprimitive group is helpful, for if N is a maximal intransitive normal subgroup of $\mathrm{Aut}(\Gamma)$ then both G and $\mathrm{Aut}(\Gamma)$ induce quasiprimitive automorphism groups of the normal quotient Γ_H, where $H = \mathrm{Aut}(\Gamma)_\alpha N$, and the action of G is faithful. This approach was used, for example, in classifying the 2-arc transitive graphs admitting $\mathrm{Sz}(q)$ or $\mathrm{Ree}(q)$ mentioned above, and also in analysing the automorphism groups of Cayley graphs of simple groups in [8].

Innately transitive groups, identified in § as a third possibility for basic groups, have not received much attention until recently. They arise naturally when investigating the full automorphism groups of graphs. One example is given in [7] for locally-primitive graphs Γ admitting an almost simple vertex-quasiprimitive subgroup G of automorphisms. It is shown that either $\mathrm{Aut}(\Gamma)$ is innately transitive, or G is of Lie type in characteristic p and $\mathrm{Aut}(\Gamma)$ has a minimal normal p-subgroup involving a known G-module.

4. Characteristics of basic permutation groups

Finite primitive permutation groups have attracted the attention of mathematicians for more than a hundred years. In particular, one of the central problems of 19th century Group Theory was to find an upper bound, much smaller than $n!$, for the order of a primitive group on a set of size n, other than the symmetric group S_n and the alternating group A_n. It is now known that the largest such groups occur for n of the form $c(c-1)/2$ and are S_c and A_c acting on the unordered pairs from a set of size c. The proofs of this and other results in this section depend on the finite simple group classification.

If G is a quasiprimitive permutation group on Ω, $\alpha \in \Omega$, and H is a maximal subgroup of G containing G_α, then the primitive component $\mathsf{Comp}(G, H)$ is isomorphic to G since the kernel of this action is an intransitive normal subgroup of G and hence is trivial. Because of this we may often deduce information about quasiprimitive groups from their primitive components, and indeed it was found in [29] that finite quasiprimitive groups possess many characteristics similar to those of finite primitive groups. This is true also of innately transitive groups. We state just one example, concerning the orders of permutation groups acting on a set of size n, that is, of *degree* n.

Theorem 4.1 [4, 29] *There is a constant c and an explicitly defined family $\mathcal{F}$ of finite permutation groups such that, if G is a primitive, quasiprimitive, or innately transitive permutation group of degree n, then either $G \in \mathcal{F}$, or $|G| < n^{c \log n}$.*

The O'Nan-Scott Theorem partitions the finite primitive permutation groups into several disjoint types according to the structure or action of their minimal normal subgroups. It highlights the role of simple groups and their representations in analysing and using primitive groups. One of its first successful applications was the analysis of distance transitive graphs in Theorem 3.1. Other early applications include a proof [6] of the Sims Conjecture, and a classification result [18] for maximal subgroups of A_n and S_n, both of which are stated below.

Theorem 4.2 [6] *There is a function f such that if G is primitive on a finite set Ω, and for $\alpha \in \Omega$, G_α has an orbit of length d in $\Omega \setminus \{\alpha\}$, then $|G_\alpha| \leq f(d)$.*

Theorem 4.3 [18] *Let $G = A_n$ or S_n with M a maximal subgroup. Then either M belongs to an explicit list or M is almost simple and primitive. Moreover if $H < G$ and H is almost simple and primitive but not maximal, then (H, n) is known.*

This is a rather curious way to state a classification result. However it seems almost inconceivable that the finite almost simple primitive groups will ever be listed explicitly. Instead [18] gives an explicit list of triples (H, M, n), where H is primitive of degree n with a nonabelian simple normal subgroup T not normalised by M, and $H < M < HA_n$. This result suggested the possibility of describing the lattice of all primitive permutation groups on a given set, for it gave a description of the over-groups of the almost simple primitive groups. Such a description was achieved in [23] using a general construction for primitive groups called a *blow-up* construction introduced by Kovacs [14]. The analysis leading to Theorem 3.4 was based on this theorem.

Theorem 4.4 [23] *All inclusions $G < H < S_n$ with G primitive are either explicitly described, or are described in terms of a blow-up of an explicitly listed inclusion $G_1 < H_1 < S_{n_1}$ with n a proper power of n_1.*

Analogues of the O'Nan-Scott Theorem for finite quasiprimitive and innately transitive groups have been proved in [3, 24] and enable similar analyses to be undertaken for problems involving these classes of groups. For example, the quasiprimitive version formed the basis for Theorems 3.2 and 3.3. It seems to be the most useful version for dealing with families of vertex-transitive or locally-transitive graphs. A description of the lattice of quasiprimitive subgroups of S_n was given in [2, 26] and was used, for example, in analysing Cayley graphs of finite simple groups in [8].

Theorem 4.5 [2, 26] *Suppose that $G < H < S_n$ with G quasiprimitive and imprimitive, and H quasiprimitive but $H \neq A_n$. Then either G and H have equal socles and the same O'Nan-Scott types, or the possibilities for the O'Nan-Scott types of G, H are restricted and are known explicitly.*

In the latter case, for most pairs of O'Nan-Scott types, explicit constructions are given for these inclusions. Not all the types of primitive groups identified by the O'Nan-Scott Theorem occur for every degree n. Let us call permutation groups of degree n other than A_n and S_n nontrivial. A systematic study by Cameron, Neumann and Teague [5] of the integers n for which there exists a nontrivial primitive group of degree n showed that the set of such integers has density zero in the natural numbers. Recently it was shown in [30] that a similar result holds for the degrees of nontrivial quasiprimitive and innately transitive permutation groups. Note that $2.2 < \sum_{d=1}^{\infty} \frac{1}{d\phi(d)} < 2.23$.

Theorem 4.6 [5, 30] *For a positive real number x, the proportion of integers $n \leq x$ for which there exists a nontrivial primitive, quasiprimitive, or innately transitive permutation group of degree n is at most $(1 + o(1))c/\log x$, where $c = 2$ in the case of primitive groups, or $c = 1 + \sum_{d=1}^{\infty} \frac{1}{d\phi(d)}$ for the other cases.*

5. Simple groups and basic permutation groups

Many of the results about basic permutation groups mentioned above rely on specific knowledge about finite simple groups. Sometimes this knowledge was already available in the simple group literature. However investigations of basic permutation groups often raised interesting new questions about simple groups. Answering these questions became an integral part of the study of basic groups, and the answers enriched our understanding of finite simple groups. In this final section we review a few of these new simple group results. Handling the primitive almost simple classical groups was the most difficult part of proving Theorem 4.3, and the following theorem of Aschbacher formed the basis for their analysis.

Theorem 5.1 [1] *Let G be a subgroup of a finite almost simple classical group X such that G does not contain* $\mathrm{soc}(X)$*, and let V denote the natural vector space associated with X. Then either G lies in one of eight explicitly defined families of subgroups, or G is almost simple, absolutely irreducible on V and the (projective) representation of* $\mathrm{soc}(G)$ *on V cannot be realised over a proper subfield.*

A detailed study of classical groups based on Theorem 5.1 led to Theorem 5.2, a classification of the maximal factorisations of the almost simple groups. This classification was fundamental to the proofs of Theorems 3.4 and 4.3, and has been used in diverse applications, for example see [9, 17].

Theorem 5.2 [19, 20] *Let G be a finite almost simple group and suppose that $G = AB$, where A, B are both maximal in G subject to not containing* $\mathrm{soc}(G)$*. Then G, A, B are explicitly listed.*

For a finite group G, let $\pi(G)$ denote the set of prime divisors of $|G|$. For many simple groups G there are small subsets of $\pi(G)$ that do not occur in the order of any proper subgroup, and it is possible to describe some of these precisely as follows.

Theorem 5.3 [21, Theorem 4, Corollaries 5 and 6] *Let G be an almost simple group with socle T, and let M be a subgroup of G not containing T.*

(a) If $G = T$ then for an explicitly defined subset $\Pi \subseteq \pi(T)$ with $|\Pi| \leq 3$, if $\Pi \subseteq \pi(M)$ then T, M are known explicitly, and in most cases $\pi(T) = \pi(M)$.
(b) If $\pi(T) \subseteq \pi(M)$ then T, M are known explicitly.

Theorem 5.3 was used in [10] to classify all innately transitive groups having no fixed-point-free elements of prime order, settling the polycirculant graph conjecture for such groups. Another application of Theorems 5.2 and 5.3 is the following factorisation theorem that was used in the proof of Theorem 4.5. It implies in particular that, if G is quasiprimitive of degree n with nonabelian and non-simple socle, then S_n and possibly A_n are the only almost simple over-groups of G.

Theorem 5.4 [26, Theorem 1.4] *Let T, S be finite nonabelian simple groups such that T has proper subgroups A, B with $T = AB$ and $A = S^\ell$ for some $\ell \geq 2$. Then $T = A_n$, $B = A_{n-1}$, where $n = |T : B|$, and A is a transitive group of degree n.*

Finally we note that Theorem 4.6 is based on the following result about indices of subgroups of finite simple groups.

Theorem 5.5 [5, 30] *For a positive real number x, the proportion of integers $n \leq x$ of the form $n = |T : M|$, where T is a nonabelian simple group and M is either a maximal subgroup or a proper subgroup, and $(T, M) \neq (A_n, A_{n-1})$, is at most $(1 + o(1))c/\log x$, where $c = 1$ or $c = \sum_{d=1}^{\infty} \frac{1}{d\phi(d)}$ respectively.*

We have presented a framework for studying finite permutation groups by identifying and analysing their basic components. The impetus for extending the theory beyond primitive groups came from the need for an appropriate theory of basic permutation groups for combinatorial applications. Developing this theory required the answers to specific questions about simple groups, and the power of the theory is largely due to its use of the finite simple group classification.

References

[1] M. Aschbacher, On the maximal subgroups of the finite classical groups, *Invent. Math.* **76** (1984), 469–514.

[2] R. Baddeley and C. E. Praeger, On primitive overgroups of quasiprimitive permutation groups, Research Report No. 2002/3, U. Western Australia, 2002.

[3] J. Bamberg and C. E. Praeger, Finite permutation groups with a transitive minimal normal subgroup, preprint, 2002.

[4] P. J. Cameron, Finite permutation groups and finite simple groups, *Bull. London Math. Soc.* **13** (1981), 1–22.

[5] P. J. Cameron, P. M. Neumann and D. N. Teague, On the degrees of primitive permutation groups, *Math. Z.* **180** (1982), 141–149.

[6] P. J. Cameron, C. E. Praeger, J. Saxl, and G. M. Seitz, On the Sims conjecture and distance transitive graphs, *Bull. London Math. Soc.* **15** (1983), 499–506.

[7] X. G. Fang, G. Havas, and C. E. Praeger, On the automorphism groups of quasiprimitive almost simple graphs, *J. Algebra* **222** (1999), 271–283.

[8] X. G. Fang, C. E. Praeger and J. Wang, On the automorphism groups of Cayley graphs of finite simple groups, *J. London Math. Soc.* (to appear).

[9] M. D. Fried, R. Guralnick and J. Saxl, Schur covers and Carlitz's conjecture, *Israel J. Math.* **82** (1993), 157–225.

[10] M. Giudici, Quasiprimitive groups with no fixed point free elements of prime order, *J. London Math. Soc.*, (to appear).

[11] M. Giudici, C. H. Li and C. E. Praeger, Analysing finite locally s-arc-transitive graphs, in preparation.

[12] A. A. Ivanov, Distance-transitive graphs and their classification, in *Investigations in algebraic theory of combinatorial objects*, Kluwer, Dordrecht, 1994, 283–378.

[13] A. A. Ivanov and C. E. Praeger, On finite affine 2-arc transitive graphs, *European J. Combin.* **14** (1993), 421–444.

[14] L. G. Kovacs, Primitive subgroups of wreath products in product action, *Proc. London Math. Soc.* (3) **58** (1989), 306–322.

[15] C. H. Li, Finite s-arc transitive graphs of prime-power order, *Bull. London Math. Soc.* **33** (2001), 129-137.
[16] C. H. Li, On finite s-arc transitive graphs of odd order, *J. Combin. Theory Ser. B* **81** (2001), 307-317.
[17] C. H. Li, The finite vertex-primitive and vertex-biprimitive s-transitive graphs for $s \geq 4$, *Trans. Amer. Math. Soc.* **353** (2001), 3511–3529.
[18] M. W. Liebeck, C. E. Praeger and J. Saxl, A classification of the maximal subgroups of the finite alternating and symmetric groups, *Proc. London Math. Soc.* **55** (1987), 299–330.
[19] M. W. Liebeck, C. E. Praeger and J. Saxl, The maximal factorisations of the finite simple groups and their automorphism groups, *Mem. Amer. Math. Soc.* No. 432, Vol. 86 (1990), 1–151.
[20] M. W. Liebeck, C. E. Praeger and J. Saxl, On factorisations of almost simple groups, *J. Algebra* **185** (1996), 409–419.
[21] M. W. Liebeck, C. E. Praeger and J. Saxl, Transitive subgroups of primitive permutation groups, *J. Algebra* **234** (2000), 291–361.
[22] M. W. Liebeck, C. E. Praeger and J. Saxl, Primitive permutation groups with a common suborbit, and edge-transitive graphs, *Proc. London Math. Soc.* (3) **84** (2002), 405–438.
[23] C. E. Praeger, The inclusion problem for finite primitive permutation groups, *Proc. London Math. Soc.* (3) **60** (1990), 68–88.
[24] C. E. Praeger, An O'Nan-Scott theorem for finite quasiprimitive permutation groups and an application to 2-arc transitive graphs, *J. London Math. Soc.* (2) 47 (1993), 227–239.
[25] C. E. Praeger, Quasiprimitive graphs. In *Surveys in combinatorics, 1997 (London)*, 65–85, Cambridge University Press, Cambridge, 1997.
[26] C. E. Praeger, Quotients and inclusions of finite quasiprimitive permutation groups, Research Report No. 2002/05, University of Western Australia, 2002.
[27] C. E. Praeger, Seminormal and subnormal subgroup lattices for transitive permutation groups, in preparation.
[28] C. E. Praeger, J. Saxl and K. Yokoyama, Distance transitive graphs and finite simple groups, *Proc. London Math. Soc.* (3) **55** (1987), 1–21.
[29] C. E. Praeger and A. Shalev, Bounds on finite quasiprimitive permutation groups, *J. Austral. Math. Soc.* **71** (2001), 243–258.
[30] C. E. Praeger and A. Shalev, Indices of subgroups of finite simple groups and quasiprimitive permutation groups, preprint, 2002.
[31] J. van Bon and A. M. Cohen, Prospective classification of distance-transitive graphs, in *Combinatorics '88 (Ravello)*, Mediterranean, Rende, 1991, 25–38.

ICM 2002 · Vol. II · 77–85

Norm Varieties and Algebraic Cobordism

Markus Rost*

Abstract

We outline briefly results and examples related with the bijectivity of the norm residue homomorphism. We define norm varieties and describe some constructions. We discuss degree formulas which form a major tool to handle norm varieties. Finally we formulate Hilbert's 90 for symbols which is the hard part of the bijectivity of the norm residue homomorphism, modulo a theorem of Voevodsky.

Introduction

This text is a brief outline of results and examples related with the bijectivity of the norm residue homomorphism—also called "Bloch-Kato conjecture" and, for the mod 2 case, "Milnor conjecture".

The starting point was a result of Voevodsky which he communicated in 1996. Voevodsky's theorem basically reduces the Bloch-Kato conjecture to the existence of norm varieties and to what I call Hilbert's 90 for symbols. Unfortunately there is no text available on Voevodsky's theorem.

In this exposition p is a prime, k is a field with $\operatorname{char} k \neq p$ and $K_n^M k$ denotes Milnor's n-th K-group of k [15], [19].

Elements in $K_n^M k/p$ of the form

$$u = \{a_1, \dots, a_n\} \bmod p$$

are called symbols (mod p, of weight n).

A field extension F of k is called a splitting field of u if $u_F = 0$ in $K_n^M F/p$.

Let

$$h_{(n,p)} \colon K_n^M k/p \to H^n_{\text{ét}}(k, \mu_p^{\otimes n}),$$
$$\{a_1, \dots, a_n\} \mapsto (a_1, \dots, a_n)$$

be the norm residue homomorphism.

*Department of Mathematics, The Ohio State University, 231 W 18th Avenue, Columbus, OH 43210, USA. E-mail: rost@math.ohio-state.edu, URL: http://www.math.ohio-state.edu/~rost

1. Norm varieties

All successful approaches to the Bloch-Kato conjecture consist of an investigation of appropriate generic splitting varieties of symbols. This goes back to the work of Merkurjev and Suslin on the case $n = 2$ who studied the K-cohomology of Severi-Brauer varieties [12]. Similarly, for the case $p = 2$ (for $n = 3$ by Merkurjev, Suslin [14] and the author [18], for all n by Voevodsky [23]) one considers certain quadrics associated with Pfister forms. For a long time it was not clear which sort of varieties one should consider for arbitrary n, p. In some cases one knew candidates, but these were non-smooth varieties and desingularizations appeared to be difficult to handle. Finally Voevodsky proposed a surprising characterization of the necessary varieties. It involves characteristic numbers and yields a beautiful relation between symbols and cobordism theory.

Definition. *Let $u = \{a_1, \dots, a_n\} \bmod p$ be a symbol. Assume that $u \neq 0$. A* norm variety *for u is a smooth proper irreducible variety X over k such that*

(1) *The function field $k(X)$ of X splits u.*
(2) $\dim X = d := p^{n-1} - 1$.
(3) $\frac{s_d(X)}{p} \not\equiv 0 \mod p$.

Here $s_d(X) \in \mathbf{Z}$ denotes the characteristic number of X given by the d-th Newton polynomial in the Chern classes of TX. It is known (by Milnor) that in dimensions $d = p^n - 1$ the number $s_d(X)$ is p-divisible for any X. If $k \subset \mathbf{C}$ one may rephrase condition (3) by saying that $X(\mathbf{C})$ is indecomposable in the complex cobordism ring mod p.

We will observe in section 2. that the conditions for a norm variety are birational invariant.

The name "norm variety" originates from some constructions of norm varieties, see section 3..

We conclude this section with the "classical" examples of norm varieties.

Example. The case $n = 2$. Assume that k contains a primitive p-th root ζ of unity. For $a, b \in k^*$ let $A_\zeta(a, b)$ be the central simple k-algebra with presentation

$$A_\zeta(a, b) = \langle u, v \mid u^p = a, v^p = b, vu = \zeta uv \rangle.$$

The Severi-Brauer variety $X(a, b)$ of $A_\zeta(a, b)$ is a norm variety for the symbol $\{a, b\} \bmod p$.

Example. The case $p = 2$. For $a_1, \dots, a_n \in k^*$ one denotes by

$$\langle\langle a_1, \dots, a_n \rangle\rangle = \bigotimes_1^n \langle 1, -a_i \rangle,$$

the associated n-fold Pfister form [9], [21]. The quadratic form

$$\varphi = \langle\langle a_1, \dots, a_{n-1} \rangle\rangle \perp \langle -a_n \rangle$$

is called a Pfister neighbor. The projective quadric $Q(\varphi)$ defined by $\varphi = 0$ is a norm variety for the symbol $\{a_1, \dots, a_n\} \bmod 2$.

2. Degree formulas

The theme of "degree formulas" goes back to Voevodsky's first text on the Milnor conjecture (although he never formulated explicitly a "formula") [22]. In this section we formulate the degree formula for the characteristic numbers s_d. It shows the birational invariance of the notion of norm varieties.

The first proof of this formula relied on Voevodsky's stable homotopy theory of algebraic varieties. Later we found a rather elementary approach [11], which is in spirit very close to "elementary" approaches to the complex cobordism ring [16], [4].

For our approach to Hilbert's 90 for symbols we use also "higher degree formulas" which again were first settled using Voevodsky's stable homotopy theory [3]. These follow meanwhile also from the "general degree formula" proved by Morel and Levine [10] in characteristic 0 using factorization theorems for birational maps [1].

We fix a prime p and a number d of the form $d = p^n - 1$.

For a proper variety X over k let

$$I(X) = \deg\big(\mathrm{CH}_0(X)\big) \subset \mathbf{Z}$$

be the image of the degree map on the group of 0-cycles. One has $I(X) = i(X)\mathbf{Z}$ where $i(X)$ is the "index" of X, i. e., the gcd of the degrees $[k(x) : k]$ of the residue class field extensions of the closed points x of X. If X has a k-point (in particular if k is algebraically closed), then $I(X) = \mathbf{Z}$. The group $I(X)$ is a birational invariant of X. We put

$$J(X) = I(X) + p\mathbf{Z}.$$

Let X, Y be irreducible smooth proper varieties over k with $\dim Y = \dim X = d$ and let $f\colon Y \to X$ be a morphism. Define $\deg f$ as follows: If $\dim f(Y) < \dim X$, then $\deg f = 0$. Otherwise $\deg f \in \mathbf{N}$ is the degree of the extension $k(Y)/k(X)$ of the function fields.

Theorem (Degree formula for s_d).

$$\frac{s_d(Y)}{p} = (\deg f)\frac{s_d(X)}{p} \mod J(X).$$

Corollary. *The class*

$$\frac{s_d(X)}{p} \bmod J(X) \in \mathbf{Z}/J(X)$$

is a birational invariant.

Remark. If X has a k-rational point, then $J(X) = \mathbf{Z}$ and the degree formula is empty. The degree formula and the birational invariants $s_d(X)/p \bmod J(X)$ are phenomena which are interesting only over non-algebraically closed fields. Over the complex numbers the only characteristic numbers which are birational invariant are the Todd numbers.

We apply the degree formula to norm varieties. Let u be a nontrivial symbol mod p and let X be a norm variety for u. Since $k(X)$ splits u, so does any residue class field $k(x)$ for $x \in X$. As u is of exponent p, it follows that $J(X) = p\mathbf{Z}$.

Corollary (Voevodsky). *Let u be a nontrivial symbol and let X be a norm variety of u. Let further Y be a smooth proper irreducible variety with $\dim Y = \dim X$ and let $f\colon Y \to X$ be a morphism. Then Y is a norm variety for u if and only if $\deg f$ is prime to p.*

It follows in particular that the notion of norm variety is birational invariant. Therefore we may call any irreducible variety U (not necessarily smooth or proper) a norm variety of a symbol u if U is birational isomorphic to a smooth and proper norm variety of u.

3. Existence of norm varieties

Theorem. *Norm varieties exists for every symbol $u \in K_n^M k/p$ for every p and every n.*

As we have noted, for the case $n = 2$ one can take appropriate Severi-Brauer varieties (if k contains the p-th roots of unity) and for the case $p = 2$ one can take appropriate quadrics.

In this exposition we describe a proof for the case $n = 3$ using fix-point theorems of Conner and Floyd in order to compute the non-triviality of the characteristic numbers. Our first proof for the general case used also Conner-Floyd fix-point theory. Later we found two further methods which are comparatively simpler. However the Conner-Floyd fix-point theorem is still used in our approach to Hilbert's 90 for symbols.

Let $u = \{a, b, c\} \bmod p$ with $a,\ b,\ c \in k^*$. Assume that k contains a primitive p-th root ζ of unity, let $A = A_\zeta(a, b)$ and let

$$MS(A, c) = \{\, x \in A \mid \mathrm{Nrd}(x) = c \,\}.$$

We call $MS(A, c)$ the Merkurjev-Suslin variety associated with A and c. The symbol u is trivial if and only if $MS(A, c)$ has a rational point [12]. The variety $MS(A, c)$ is a twisted form of $\mathrm{SL}(p)$.

Theorem. *Suppose $u \neq 0$. Then $MS(A, c)$ is a norm variety for u.*

Let us indicate a proof for a subfield $k \subset \mathbf{C}$ (and for $p > 2$). Let $U = MS(A, c)$. It is easy to see that $k(U)$ splits u. Moreover one has $\dim U = \dim A - 1 = p^2 - 1$. It remains to show that there exists a proper smooth completion X of U with nontrivial characteristic number.

Let

$$\bar{U} = \{\, [x, t] \in \mathbf{P}(A \oplus k) \mid \mathrm{Nrd}(x) = ct^p \,\}$$

be the naive completion of U. We let the group $G = \mathbf{Z}/p \times \mathbf{Z}/p$ act on the algebra A via

$$(r, s) \cdot u = \zeta^r u, \quad (r, s) \cdot v = \zeta^s v.$$

This action extends to an action on $\mathbf{P}(A \oplus k)$ (with the trivial action on k) which induces a G-action on $\bar{U}$. Let $\mathrm{Fix}(\bar{U})$ be the fixed point scheme of this action. One

finds that $\mathrm{Fix}(\bar{U})$ consists just of the p isolated points $[1, \zeta^i]$, $i = 1, \ldots, p$, which are all contained in U.

The variety U is smooth, but $\bar{U}$ is not. However, by equivariant resolution of singularities [2], there exists a smooth proper G-variety X together with a G-morphism $X \to \bar{U}$ which is a birational isomorphism and an isomorphism over U. It remains to show that

$$\frac{s_d(X)}{p} \not\equiv 0 \mod p.$$

For this we may pass to topology and try to compute $s_d\big(X(\mathbf{C})\big)$. We note that for odd p, the Chern number s_d is also a Pontryagin number and depends only on the differentiable structure of the given variety. Note further that X has the same G-fixed points as $\bar{U}$ since the desingularization took place only outside U.

Consider the variety

$$Z = \left\{ \left[\textstyle\sum_{i,j=1}^{p} x_{ij} u^i v^j, t \right] \in \mathbf{P}(A \oplus k) \;\middle|\; \textstyle\sum_{i,j=1}^{p} x_{ij}^p = t^p \right\}.$$

This variety is a smooth hypersurface and it is easy to check

$$\frac{s_d(Z)}{p} \not\equiv 0 \mod p.$$

As a G-variety, the variety Z has the same fixed points as X ("same" means that the collections of fix-points together with the G-structure on the tangent spaces are isomorphic). Let M be the differentiable manifold obtained from $X(\mathbf{C})$ and $-Z(\mathbf{C})$ by a multi-fold connected sum along corresponding fixed points. Then M is a G-manifold without fixed points. By the theory of Conner and Floyd [5], [7] applied to $(\mathbf{Z}/p)^2$-manifolds of dimension $d = p^2 - 1$ one has

$$\frac{s_d(M)}{p} \equiv 0 \mod p.$$

Thus

$$\frac{s_d(X)}{p} \equiv \frac{s_d(Z)}{p} \mod p$$

and the desired non-triviality is established.

The functions Φ_n. We conclude this section with examples of norm varieties for the general case.

Let $a_1, a_2, \ldots$ be a sequence of elements in k^*. We define functions $\Phi_n = \Phi_{a_1,\ldots,a_n}$ in p^n variables inductively as follows.

$$\Phi_0(t) = t^p,$$

$$\Phi_n(T_0, \ldots, T_{p-1}) = \Phi_{n-1}(T_0) \prod_{i=1}^{p-1} \big(1 - a_n \Phi_{n-1}(T_i)\big).$$

Here the T_i stand for tuples of p^{n-1} variables. Let $U(a_1, \ldots, a_n)$ be the variety defined by

$$\Phi_{a_1,\ldots,a_{n-1}}(T) = a_n.$$

Theorem. *Suppose that the symbol $u = \{a_1, \ldots, a_n\} \bmod p$ is nontrivial. Then $U(a_1, \ldots, a_n)$ is a norm variety of u.*

4. Hilbert's 90 for symbols

The bijectivity of the norm residue homomorphisms has always been considered as a sort of higher version of the classical Hilbert's Theorem 90 (which establishes the bijectivity for $n = 1$). In fact, there are various variants of the Bloch-Kato conjecture which are obvious generalizations of Hilbert's Theorem 90: The Hilbert's Theorem 90 for K_n^M of cyclic extensions or the vanishing of the motivic cohomology group $H^{n+1}\big(k, \mathbf{Z}(n)\big)$. In this section we describe a variant which on one hand is very elementary to formulate and on the other hand is the really hard part of the Bloch-Kato conjecture (modulo Voevodsky's theorem).

Let $u = \{a_1, \dots, a_n\} \in K_n^M k/p$ be a symbol. Consider the norm map

$$\mathcal{N}_u = \sum_F N_{F/k} \colon \bigoplus_F K_1 F \to K_1 k$$

where F runs through the finite field extensions of k (contained in some algebraic closure of k) which split u. Hilbert's Theorem 90 for u states that $\ker \mathcal{N}_u$ is generated by the "obvious" elements.

To make this precise, we consider two types of basic relations between the norm maps $N_{F/k}$.

Let F_1, F_2 be finite field extensions of k. Then the sequence

$$K_1(F_1 \otimes F_2) \xrightarrow{(N_{F_1 \otimes F_2/F_1}, -N_{F_1 \otimes F_2/F_2})} K_1 F_1 \oplus K_1 F_2 \xrightarrow{N_{F_1/k} + N_{F_2/k}} K_1 k \qquad (1)$$

is a complex.

Further, if K/k is of transcendence degree 1, then the sequence

$$K_2 K \xrightarrow{d_K} \bigoplus_v K_1 \kappa(v) \xrightarrow{N} K_1 k \qquad (2)$$

is a complex. Here v runs through the valuations of K/k, d_K is given by the tame symbols at each v and N is the sum of the norm maps $N_{\kappa(v)/k}$. The sum formula $N \circ d_K = 0$ is also known as Weil's formula.

We now restrict again to splitting fields of u. The maps in (1) yield a map

$$\mathcal{R}_u = \sum_{F_1, F_2} (N_{F_1 \otimes F_2/F_1}, -N_{F_1 \otimes F_2/F_2}) \colon \bigoplus_{F_1, F_2} K_1(F_1 \otimes F_2) \to \bigoplus_F K_1 F$$

with $\mathcal{N}_u \circ \mathcal{R}_u = 0$. Let C be the cokernel of $\mathcal{R}_u$ and let $\mathcal{N}'_u \colon C \to K_1 k$ be the map induced by $\mathcal{N}_u$. Then the maps in (2) yield a map

$$\mathcal{S}_u = \sum_K d_K \colon \bigoplus_K K_2 K \to C$$

with $\mathcal{N}'_u \circ \mathcal{S}_u = 0$ where K runs through the splitting fields of u of transcendence degree 1 over k (contained in some universal field). Let $H_0(u, K_1)$ be the cokernel of $\mathcal{S}_u$ and let $N_u \colon H_0(u, K_1) \to K_1 k$ be the map induced by $\mathcal{N}'_u$.

Hilbert's 90 for symbols. *For every symbol u the norm map*

$$N_u\colon H_0(u, K_1) \to K_1 k$$

is injective.

Example. If $u = 0$, then it is easy to see that N_u is injective. In fact, it is a trivial exercise to check that $\mathcal{N}'_u$ is injective.

Example. The case $n = 1$. The splitting fields F of $u = \{a\} \bmod p$ are exactly the field extensions of k containing a p-th root of a. It is an easy exercise to reduce the injectivity of N_u (in fact of $\mathcal{N}'_u$) to the classical Hilbert's Theorem 90, i. e., the exactness of

$$K_1 L \xrightarrow{1-\sigma} K_1 L \xrightarrow{N_{L/k}} K_1 k$$

for a cyclic extension L/k of degree p with σ a generator of $\mathrm{Gal}(L/k)$.

Example. The case $n = 2$. Assume that k contains a primitive p-th root ζ of unity. The splitting fields F of $u = \{a, b\} \bmod p$ are exactly the splitting fields of the algebra $A_\zeta(a, b)$. One can show that

$$H_0(u, K_1) = K_1 A_\zeta(a, b)$$

with N_u corresponding to the reduced norm map Nrd [13]. Hence in this case Hilbert's 90 for u reduces to the classical fact $SK_1 A = 0$ for central simple algebras of prime degree [6].

Example. The case $p = 2$. The splitting fields F of $u = \{a_1, \ldots, a_n\} \bmod 2$ are exactly the field extensions of k which split the Pfister form $\langle\langle a_1, \ldots, a_n \rangle\rangle$ or, equivalently, over which the Pfister neighbor $\langle\langle a_1, \ldots, a_{n-1} \rangle\rangle \perp \langle -a_n \rangle$ becomes isotropic. Hilbert's 90 for symbols mod 2 had been first established in [17]. This text considered similar norm maps associated with any quadratic form (which are not injective in general). A treatment of the special case of Pfister forms is contained in [8].

Remark. One can show that the group $H_0(u, K_1)$ as defined above is also the quotient of $\oplus_F K_1 F$ by the R-trivial elements in $\ker \mathcal{N}_u$. This is quite analogous to the description of $K_1 A$ of a central simple algebra A: The group $K_1 A$ is the quotient of A^* by the subgroup of R-trivial elements in the kernel of Nrd: $A^* \to F^*$. Similarly for the case $p = 2$: In this case the injectivity of N_u is related with the fact that for Pfister neighbors φ the kernel of the spinor norm $\mathrm{SO}(\varphi) \to k^*/(k^*)^2$ is R-trivial.

In our approach to Hilbert's 90 for symbols one needs a parameterization of the splitting fields of symbols.

Definition. *Let $u = \{a_1, \ldots, a_n\} \bmod p$ be a symbol. A* p-generic splitting variety *for u is a smooth variety X over k such that for every splitting field F of u there exists a finite extension F'/F of degree prime to p and a morphism* $\mathrm{Spec}\, F' \to X$.

Theorem. *Suppose* $\mathrm{char}\, k = 0$. *Let $m \geq 3$ and suppose for $n \leq m$ and every symbol $u = \{a_1, \ldots, a_n\} \bmod p$ over all fields over k there exists a p-generic splitting variety for u of dimension $p^{n-1} - 1$. Then Hilbert's 90 holds for such symbols.*

The proof of this theorem is outlined in [20].

For $n = 2$ one can take here the Severi-Brauer varieties and for $n = 3$ the Merkurjev-Suslin varieties. Hence we have:

Corollary. *Suppose* $\operatorname{char} k = 0$. *Then Hilbert's 90 holds for symbols of weight* ≤ 3.

References

[1] D. Abramovich, K. Karu, K. Matsuki, and J. Włodarczyk, *Torification and factorization of birational maps*, J. Amer. Math. Soc. **15** (2002), no. 3, 531–572 (electronic).

[2] E. Bierstone and P. D. Milman, *Canonical desingularization in characteristic zero by blowing up the maximum strata of a local invariant*, Invent. Math. **128** (1997), no. 2, 207–302.

[3] S. Borghesi, *Algebraic Morava K-theories and the higher degree formula*, thesis, 2000, Evanston, http://www.math.uiuc.edu/K-theory/0412

[4] S. Buoncristiano and D. Hacon, *The geometry of Chern numbers*, Ann. of Math. (2) **118** (1983), no. 1, 1–7.

[5] P. E. Conner and E. E. Floyd, *Differentiable periodic maps*, Academic Press Inc., Publishers, New York, 1964, Ergebnisse der Mathematik und ihrer Grenzgebiete, N. F., Band 33.

[6] P. K. Draxl, *Skew fields*, London Mathematical Society Lecture Note Series, vol. 81, Cambridge University Press, Cambridge, 1983.

[7] E. E. Floyd, *Actions of* $(Z_p)^k$ *without stationary points*, Topology **10** (1971), 327–336.

[8] B. Kahn, *La conjecture de Milnor (d'après V. Voevodsky)*, Astérisque (1997), no. 245, Exp. No. 834, 5, 379–418, Séminaire Bourbaki, Vol. 1996/97.

[9] T. Y. Lam, *The algebraic theory of quadratic forms*, Benjamin/Cummings Publishing Co. Inc. Advanced Book Program, Reading, Mass., 1980, Revised second printing, Mathematics Lecture Note Series.

[10] M. Levine and F. Morel, *Algebraic cobordism I*, preprint, 2002, http://www.math.uiuc.edu/K-theory/0547

[11] A. S. Merkurjev, *Degree formula*, notes, May 2000, http://www.math.ohio-state.edu/~rost/chain-lemma.html

[12] A. S. Merkurjev and A. A. Suslin, *K-cohomology of Severi-Brauer varieties and the norm residue homomorphism*, Izv. Akad. Nauk SSSR Ser. Mat. **46** (1982), no. 5, 1011–1046, 1135–1136 (Russian), [Math. USSR Izv. **21** (1983), 307–340].

[13] ———, *The group of* K_1*-zero-cycles on Severi-Brauer varieties*, Nova J. Algebra Geom. **1** (1992), no. 3, 297–315.

[14] ———, *Norm residue homomorphism of degree three*, Izv. Akad. Nauk SSSR Ser. Mat. **54** (1990), no. 2, 339–356 (Russian), [Math. USSR Izv. **36** (1991), no. 2, 349–367], also: LOMI-preprint (1986).

[15] J. Milnor, *Algebraic K-theory and quadratic forms*, Invent. Math. **9** (1970), 318–344.

[16] D. Quillen, *Elementary proofs of some results of cobordism theory using Steen-*

rod operations, Advances in Math. **7** (1971), 29–56 (1971).
[17] M. Rost, *On the spinor norm and $A_0(X, K_1)$ for quadrics*, preprint, 1988, http://www.math.ohio-state.edu/~rost/spinor.html
[18] ———, *Hilbert 90 for K_3 for degree-two extensions*, Preprint, 1986.
[19] ———, *Chow groups with coefficients*, Doc. Math. **1** (1996), No. 16, 319–393 (electronic).
[20] ———, *Chain lemma for splitting fields of symbols*, preprint, 1998, http://www.math.ohio-state.edu/~rost/chain-lemma.html
[21] W. Scharlau, *Quadratic and Hermitian forms*, Grundlehren der mathematischen Wissenschaften, vol. 270, Springer-Verlag, Berlin, 1985.
[22] V. Voevodsky, *The Milnor conjecture*, preprint, 1996, Max-Planck-Institute for Mathematics, Bonn, http://www.math.uiuc.edu/K-theory/0170/
[23] ———, *On 2-torsion in motivic cohomology*, preprint, 2001, http://www.math.uiuc.edu/K-theory/0502/

ICM 2002 · Vol. II · 87–92

Diophantine Geometry over Groups and the Elementary Theory of Free and Hyperbolic Groups*

Z. Sela†

Abstract

We study sets of solutions to equations over a free group, projections of such sets, and the structure of elementary sets defined over a free group. The structre theory we obtain enable us to answer some questions of A. Tarski's, and classify those finitely generated groups that are elementary equivalent to a free group. Connections with low dimensional topology, and a generalization to (Gromov) hyperbolic groups will also be discussed.

2000 Mathematics Subject Classification: 14, 20.

Sets of solutions to equations defined over a free group have been studied extensively, mostly since Alfred Tarski presented his fundamental questions on the elementary theory of free groups in the mid 1940's. Considerable progress in the study of such sets of solutions was made by G. S. Makanin, who constructed an algorithm that decides if a system of equations defined over a free group has a solution [Ma1], and showed that the universal and positive theories of a free group are decidable [Ma2]. A. A. Razborov was able to give a description of the entire set of solutions to a system of equations defined over a free group [Ra], a description that was further developed by O. Kharlampovich and A. Myasnikov [Kh-My].

A set of solutions to equations defined over a free group is clearly a discrete set, and all the previous techniques and methods that studied these sets are combinatorial in nature. Naturally, the structure of sets of solutions defined over a free group is very different from the structure of sets of solutions (varieties) to systems of equations defined over the complexes, reals or a number field. Still, perhaps surprisingly, concepts from complex algebraic geometry and from Diophantine geometry can be borrowed to study varieties defined over a free group.

*Partially supported by an Israel academy of sciences fellowship, an NSF grant DMS9729992 through the IAS, and the IHES.

†Hebrew University, Jerusalem 91904, Israel. E-mail: zlil@math.huji.ac.il

In this work we borrow concepts and techniques from geometric group theory, low dimensional topology, and Diophantine geometry to study the structure of varieties defined over a free (and hyperbolic) group. Our techniques and point of view on the study of these varieties is rather different from any of the pre-existing techniques in this field, though, as one can expect, some of our preliminary results overlap with previously known ones. The techniques and concepts we use enable the study of the structure of varieties defined over a free group and their projections (Diophantine sets), and in particular, give us the possibility to answer some questions that seem to be essential in any attempt to understand the structure of elementary sentences and predicates defined over a free (and hyperbolic) group.

In this note we summarize the main results of our work, that enable one to answer affirmatively some of A. Tarski's problems on the elementary theory of a free group, and classify those finitely generated groups that are elementary equivalent to a (non-abelian) free group. we further survey some of our results on the elementary theory of a (torsion-free) hyperbolic group, that generalize the results on free groups. The work itself appears in [Se1]-[Se8].

We start with what we see as the main result on the elementary theory of a free group we obtained - quantifier elimination. Quantifier elimination and its proof is behind all the other results presented in this note.

Theorem 1 ([Se7],1). *Let F be a non-abelian free group, and let $Q(p)$ be a definable set over F. Then $Q(p)$ is in the Boolean algebra of AE sets over F.*

In fact it is possible to give a strengthening of theorem 1 that specifies a subclass of AE sets that generates the Boolean algebra of definable sets, a more refined description that is essential in studying other model-theoretic properties of the elementary theory of a free group.

Theorem 1 proves that every definable set over a free group is in the Boolean algebra of AE sets. To answer Tarski's questions on the elementary theory of a free group, i.e., to show the equivalence of the elementary theories of free groups of various ranks, we need to show that for coefficient free predicates, our quantifier elimination procedure does not depend on the rank of the coefficient group.

Theorem 2 ([Se7],2). *Let $Q(p)$ be a set defined by a coefficient-free predicate over a group. Then there exists a set $L(p)$ defined by a coefficient-free predicate which is in the Boolean algebra of AE predicates, so that for every non-abelian free group F, the sets $Q(p)$ and $L(p)$ are equivalent.*

Theorem 2 proves that in handling coefficient-free predicates, our quantifier elimination procedure does not depend on the rank of the coefficient (free) group. This together with the equivalence of the AE theories of free groups ([Sa],[Hr]) implies an affirmative answer to Tarski's problem on the equivalence of the elementary theories of free groups.

Theorem 3 ([Se7],3). *The elementary theories of non-abelian free groups are equivalent.*

Arguments similar to the ones used to prove theorems 2 and 3, enable us to answer affirmatively another question of Tarski's.

Theorem 4 ([Se7],4). *Let F_k, F_ℓ be free groups for $2 \leq k \leq \ell$. Then the standard embedding $F_k \to F_\ell$ is an elementary embedding.*

More generally, let F, F_1 be non-abelian free groups, let F_2 be a free group, and suppose that $F = F_1 * F_2$. Then the standard embedding $F_1 \to F$ is an elementary embedding.

Tarski's problems deal with the equivalence of the elementary theories of free groups of different ranks. Our next goal is to get a classification of all the f.g. groups that are elementary equivalent to a free group.

Non-abelian ω-residually free groups (limit groups) are known to be the f.g. groups that are universally equivalent to a non-abelian free group. If a limit group contains a free abelian group of rank 2, it can not be elementary equivalent to a free group. Hence, a f.g. group that is elementary equivalent to a non-abelian free group must be a non-elementary (Gromov) hyperbolic limit group. However, not every non-elementary hyperbolic limit group is elementary equivalent to a free group.
To demonstrate that we look at the following example. Suppose that $G = F *_{<w>} F =< b_1, b_2 > *_{<w>} < b_3, b_4 >$ is a double of a free group of rank 2, suppose that w has no roots in F, and suppose that the given amalgamated product is the abelian JSJ decomposition of the group G. By our assumptions, G is a hyperbolic limit group (see [Se1], theorem 5.12).

Claim 5 ([Se7],5). *The group $G = F *_{<w>} F$ is not elementary equivalent to the free group F.*

In section 6 of [Se1] we have presented ω-residually free towers, as an example of limit groups (the same groups are presented in [Kh-My] as well, and are called there NTQ groups).

A hyperbolic ω-residually free tower is constructed in finitely many steps. In its first level there is a non-cyclic free product of (possibly none) (closed) surface groups and a (possibly trivial) free group, where each surface in this free product is a hyperbolic surface (i.e., with negative Euler characteristic), except the non-orientable surface of genus 2. In each additional level we add a punctured surface that is amalgamated to the group associated with the previous levels along its boundary components, and in addition there exists a retract map of the obtained group onto the group associated with the previous levels. The punctured surfaces are supposed to be of Euler characteristic bounded above by -2, or a punctured torus.

The procedure used for eliminating quantifiers over a free group enables us to show that every hyperbolic ω-residually free tower is elementary equivalent to a free group. The converse is obtained by using basic properties of the JSJ decomposition and the (canonical) Makanin-Razborov diagram of a limit group ([Se7], theorem 6).

Therefore, we are finally able to get a classification of those f.g. groups that are elementary equivalent to a free group.

Theorem 6 ([Se7],7). *A f.g. group is elementary equivalent to a non-abelian free group if and only if it is a non-elementary hyperbolic ω-residually free tower.*

So far we summarized the main results of our work, that enable one to answer affirmatively some of A. Tarski's problems on the elementary theory of a free group, and classify those finitely generated groups that are elementary equivalent to a (non-abelian) free group. In the rest of this note we survey some of our results on the elementary theory of a (torsion-free) hyperbolic group, that generalize the results presented for a free group.

In the case of a free group, we have shown that every definable set is in the Boolean algebra of AE sets. The same holds for a general hyperbolic group.

Theorem 7 ([Se8],6.5). *Let Γ be a non-elementary torsion-free hyperbolic group, and let $Q(p)$ be a definable set over Γ. Then $Q(p)$ is in the Boolean algebra of AE sets over Γ.*

Furthermore, if $Q(p)$ is a set defined by a coefficient-free predicate defined over Γ, then $Q(p)$ can be defined by a coefficient-free predicate which is in the Boolean algebra of AE predicates.

The procedure used for quantifier elimination over a free group enabled us to get a classification of those f.g. groups that are elementary equivalent to a free group (theorem 6). In a similar way, it is possible to get a classification of those f.g. groups that are elementary equivalent to a given torsion-free hyperbolic group.
We start with the following basic fact, that shows the elementary invariance of negative curvature in groups.

Theorem 8 ([Se8],7.10). *Let Γ be a torsion-free hyperbolic group, and let G be a f.g. group. If G is elementary equivalent to Γ, then G is a torsion-free hyperbolic group.*

Theorem 8 restricts the class of f.g. groups that are elementary equivalent to a given hyperbolic group, to the class of hyperbolic groups. To present the elementary classification of hyperbolic groups we start with the following basic fact.

Proposition 9 ([Se8],7.1). *Let Γ_1, Γ_2 be non-elementary torsion-free rigid hyperbolic groups (i.e., Γ_1 and Γ_2 are freely-indecomposable and do not admit any non-trivial cyclic splitting). Then Γ_1 is elementary equivalent to Γ_2 if and only if Γ_1 is isomorphic to Γ_2.*

Proposition 9 implies that, in particular, a uniform lattice in a real rank 1 semi-simple Lie group that is not $SL_2(R)$ is elementary equivalent to another such lattice if and only if the two lattices are isomorphic, hence, by Mostow's rigidity the two lattices are conjugate in the same Lie group. By Margulis's normality and super-rigidity theorems, the same hold in higher rank (real) Lie groups.

Theorem 10 ([Se8],7.2). *Let L_1, L_2 be uniform lattices in real semi-simple Lie groups that are not $SL_2(R)$. Then L_1 is elementary equivalent to L_2 if and only if L_1 and L_2 are conjugate lattices in the same real Lie group G.*

Proposition 9 shows that rigid hyperbolic groups are elementary equivalent if and only if they are isomorphic. To classify elementary equivalence classes of hyperbolic groups in general, we associate with every (torsion-free) hyperbolic group Γ, a subgroup of it, that we call the *elementary core* of Γ, and denote $EC(\Gamma)$. The elementary core is a retract of the ambient hyperbolic group Γ, and although it is not canonical, its isomorphism type is an invariant of the ambient hyperbolic group. The elementary core is constructed iteratively from the ambient hyperbolic group as we describe in definition 7.5 in [Se8].
The elementary core of a hyperbolic group is a prototype for its elementary theory.

Theorem 11 ([Se8],7.6). *Let Γ be a non-elementary torsion-free hyperbolic group that is not a ω-residually free tower, i.e., that is not elementary equivalent to a free group. Then Γ is elementary equivalent to its elementary core $EC(\Gamma)$. Furthermore, the embedding of the elementary core $EC(\Gamma)$ in the ambient group Γ is an elementary embedding.*

Finally, the elementary core is a complete invariant of the class of groups that are elementary equivalent to a given (torsion-free) hyperbolic group.

Theorem 12 ([Se8],7.9). *Let Γ_1, Γ_2 be two non-elementary torsion-free hyperbolic groups. Then Γ_1 and Γ_2 are elementary equivalent if and only if their elementary cores $EC(\Gamma_1)$ and $EC(\Gamma_2)$ are isomorphic.*

Theorem 12 asserts that the elementary class of a torsion-free hyperbolic group is determined by the isomorphism class of its elementary core. Hence, in order to be able to decide whether two torsion-free hyperbolic groups are elementary equivalent one needs to compute their elementary core, and to decide if the two elementary cores are isomorphic. Both can be done using the solution to the isomorphism problem for torsion-free hyperbolic groups.

Theorem 13 ([Se8],7.11). *Let $\Gamma_1, \Gamma 2$ be two torsion-free hyperbolic groups. Then it is decidable if Γ_1 is elementary equivalent to Γ_2.*

References

[Hr] E. Hrushovski, *private communication.*

[Kh-My] O. Kharlampovich and A. Myasnikov, *Irreducible affine varieties over a free group II*, Jour. of Algebra **200** (1998), 517–570.

[Ma1] G. S. Makanin, *Equations in a free group*, Math. USSR Izvestiya **21** (1983), 449–469.

[Ma2] ———, *Decidability of the universal and positive theories of a free group*, Math. USSR Izvestiya **25** (1985), 75–88.

[Ra] A. A. Razborov, *On systems of equations in a free group*, Ph.D. thesis, Steklov Math. institute, 1987.

[Sa] G. S. Sacerdote, *Elementary properties of free groups*, Transactions Amer. Math. Soc. **178** (1973), 127–138.

[Se1-Se8] Z. Sela, *Diophantine geometry over groups I-VIII*, preprints, www.ma.huji.ac.il/ zlil.

ICM 2002 · Vol. II · 93–103

Noncommutative Projective Geometry*

J. T. Stafford†

Abstract

This article describes recent applications of algebraic geometry to noncommutative algebra. These techniques have been particularly successful in describing graded algebras of small dimension.

2000 Mathematics Subject Classification: 14A22, 16P40, 16W50.
Keywords and Phrases: Noncommutative projective geometry, Noetherian graded rings, Deformations, Twisted homogeneous coordinate rings.

1. Introduction

In recent years a surprising number of significant insights and results in noncommutative algebra have been obtained by using the global techniques of projective algebraic geometry. This article will survey some of these results.

The classical approach to projective geometry, where one relates a commutative graded domain C to the associated variety $X = \operatorname{Proj} C$ of homogeneous, nonirrelevant prime ideals, does not generalize well to the noncommutative situation, simply because noncommutative algebras do not have enough ideals. However, there is a second approach, based on a classic theorem of Serre: If C is generated in degree one, then the categories $\operatorname{coh}(X)$ of coherent sheaves on X and $\operatorname{qgr} C$ of finitely generated graded C-modules modulo torsion are equivalent.

Surprisingly, noncommutative analogues of this idea work very well and have lead to a number of deep results. There are two strands to this approach. First, since X can be reconstructed from $\operatorname{coh}(X)$ [21] we will regard $\operatorname{coh}(X)$ rather than X as the variety since this is what generalizes. Thus, given a noncommutative graded k-algebra $R = \bigoplus R_i$ generated in degree one we will consider $\operatorname{qgr} R$ as the corresponding "noncommutative variety" (the formal definitions will be given in a moment). In particular, we will regard $\operatorname{qgr} R$ as a noncommutative curve, respectively surface, if $\dim_k R_i$ grows linearly, respectively quadratically. This analogy works well, since there are many situations in which one can pass back and forth

*The author is supported in part by the National Science Foundation under grant DMS-9801148.
†Department of Mathematics, The University of Michigan, Ann Arbor, MI 48109-1109, USA. E-mail: jts@umich.edu

between R and $\operatorname{qgr} R$ [8] and, moreover, substantial geometric techniques can be applied to study $\operatorname{qgr} R$. A survey of this approach may be found in [25].

The second strand is more concrete. In order to use algebraic geometry to study noncommutative algebras we need to be able to create honest varieties from those algebras. This is frequently possible and such an approach will form the basis of this survey. Once again, the idea is simple: when R is commutative, the points of $\operatorname{Proj} R$ correspond to the graded factor modules $M = R/I = \bigoplus_{i \geq 0} M_i$ for which $\dim_k M_i = 1$ for all i. These modules are still defined when R is noncommutative and are called point modules. In many circumstances the set of all such modules is parametrized by a commutative scheme and that scheme controls the structure of R.

This article surveys significant applications of this idea. Notably:

- If $R = \bigoplus R_i$ is a domain such that $\dim_k R_i$ grows linearly, then $\operatorname{qgr} R \simeq \operatorname{coh}(X)$ for a curve X and R can be reconstructed from data on X. Thus, noncommutative curves are commutative (see Section 4).
- The noncommutative analogues $\operatorname{qgr} R$ of the projective plane can be classified. In this case, the point modules are parametrized by either $\mathbb{P}^2$ (in which case $\operatorname{qgr} R \simeq \mathbb{P}^2$) or by an cubic curve $E \subset \mathbb{P}^2$, in which case data on E determines R (see Section 2).
- For strongly noetherian rings, as defined in Section 5, the point modules are always parametrized by a projective scheme. However there exist many noetherian algebras R for which no such parametrization exists. This has interesting consequences for the classification of noncommutative surfaces.

We now make precise the definitions that will hold throughout this article. All rings will be algebras over a fixed, algebraically closed base field k (although most of the results actually hold for arbitrary fields). A k-algebra R is called *connected graded* (*cg*) if R is a finitely generated $\mathbb{N}$-graded k-algebra $R = \bigoplus_{i \geq 0} R_i$ with $R_0 = k$. Note that this forces $\dim_k R_i < \infty$ for all i. Usually, we will assume that R is generated in degree one in the sense that R is generated by R_1 as a k-algebra. If $R = \bigoplus_{i \in \mathbb{N}} R_i$ is a right noetherian cg ring then define $\operatorname{gr} R$ to be the category of finitely generated, $\mathbb{Z}$-graded right R-modules, with morphisms being graded homomorphisms of degree zero. Define the torsion subcategory, $\operatorname{tors} R$, to be the full subcategory of $\operatorname{gr} R$ generated by the finite dimensional modules and write $\operatorname{qgr} R = \operatorname{gr} R / \operatorname{tors} R$. We write π for the canonical morphism $\operatorname{gr} R \to \operatorname{qgr} R$ and set $\mathcal{R} = \pi(R)$.

One can—and often should—work more generally with all graded R-modules and all quasi-coherent sheaves of $\mathcal{O}_X$-modules, but two categories are enough.

In order to measure the growth of an algebra we use the following dimension function: For a cg ring $R = \bigoplus_{i \geq 0} R_i$, the *Gelfand-Kirillov dimension* of R is defined to be $\operatorname{GKdim} R = \inf \{\alpha \in \mathbb{R} : \dim_k (\sum_{i=0}^{n} R_i) \leq n^\alpha \text{ for all } n \gg 0\}$. Basic facts about this dimension can be found in [17]. If R is a commutative cg algebra then $\operatorname{GKdim} R$ equals the Krull dimension of R and hence equals $\dim \operatorname{Proj} R + 1$. Thus a noncommutative curve, respectively surface, will more formally be defined as $\operatorname{qgr} R$ for a cg algebra R with $\operatorname{GKdim} R = 2$, respectively 3.

2. Historical background

We begin with a historical introduction to the subject. It really started with the work of Artin and Schelter [2] who attempted to classify the noncommutative analogues R of a polynomial ring in three variables (and therefore of $\mathbb{P}^2$). The first problem is one of definition. A "noncommutative polynomial ring" should obviously be a cg ring of finite global dimension, but this is too general, since it includes the free algebra. One can circumvent this problem by requiring that $\dim_k R_i$ grows polynomially, but this still does not exclude unpleasant rings like $k\{x,y\}/(xy)$ that has global dimension two but is neither noetherian nor a domain. The solution is to impose a Gorenstein condition and this leads to the following definition:

Definition 1 *A cg algebra R is called* AS-regular of dimension d *if* $\operatorname{gl\,dim} R = d$, $\operatorname{GKdim} R < \infty$ *and R is AS-Gorenstein; that is,* $\operatorname{Ext}^i(k,R) = 0$ *for $i \neq d$ but* $\operatorname{Ext}^d(k,R) = k$, *up to a shift of degree.*

One advantage with the Gorenstein hypothesis, for AS-regular rings of dimension 3, is that the projective resolution of k is forced to be of the form

$$0 \longrightarrow R \longrightarrow R^{(n)} \longrightarrow R^{(n)} \longrightarrow R \longrightarrow k \longrightarrow 0$$

for some n and, as Artin and Schelter show in [2], this gives strong information on the Hilbert series and hence the defining relations of R. In the process they constructed one class of algebras that they were unable to analyse:

Example 2 The three-dimensional Sklyanin algebra is the algebra

$$\mathrm{Skl}_3 = \mathrm{Skl}_3(a,b,c) = k\{x_0,x_1,x_2\}/(ax_ix_{i+1} + bx_{i+1}x_i + cx_{i+2}^2 : i \in \mathbb{Z}_3),$$

where $(a,b,c) \in \mathbb{P}^2 \setminus F$, for a (known) set F.

The original Sklyanin algebra Skl_4 is a 4-dimensional analogue of Skl_3 discovered in [23]. Independently of [2], Odesskii and Feigin [18] constructed analogues of Skl_4 in all dimensions and coined the name Sklyanin algebra. See [13] for applications of Sklyanin algebras to another version of noncommutative geometry.

In retrospect the reason Skl_3 is hard to analyse is because it depends upon an elliptic curve and so a more geometric approach is required. This approach came in [6] and depended upon the following simple idea. Assume that R is a cg algebra that is generated in degree one. Define a *point module* to be a cyclic graded (right) R-module $M = \bigoplus_{i\geq 0} M_i$ such that $\dim_k M_i = 1$ for all $i \geq 0$. The notation is justified by the fact that, if R were commutative, then such a point module M would be isomorphic to $k[x]$ and hence equal to the homogeneous coordinate ring of a point in $\operatorname{Proj} R$. Point modules are easy to analyse geometrically and this provides an avenue for using geometry in the study of cg rings.

We will illustrate this approach for $S = \mathrm{Skl}_3$. Given a point module $M = \bigoplus M_i$ write $M_i = m_i k$ for some $m_i \in M_i$ and suppose that the module structure is defined by $m_i x_j = \lambda_{ij} m_{i+1}$ for some $\lambda_{ij} \in k$. If $f = \sum f_{ij} x_i x_j$ is one of the relations for S, then necessarily $m_0 f = (\sum f_{ij}\lambda_{0i}\lambda_{1j})m_2$, whence $\sum f_{ij}\lambda_{0i}\lambda_{1j} = 0$.

This defines a subvariety $\Gamma \subseteq \mathbb{P}(S_1^*) \times \mathbb{P}(S_1^*) = \mathbb{P}^2 \times \mathbb{P}^2$ and clearly Γ parametrizes the *truncated point modules of length three*: cyclic R-modules $M = M_0 \oplus M_1 \oplus M_2$ with $\dim M_i = 1$ for $0 \leq i \leq 2$. A simple computation (see [6, Section 3] or [25, Section 8]) shows that Γ is actually the graph of an automorphism σ of an elliptic curve $E \subset \mathbb{P}^2$. It follows easily that Γ also parametrizes the point modules. As a morphism of point modules, σ is nothing more than the shift functor $M = \bigoplus M_i \mapsto M_{\geq 1}[1] = M_1 \oplus M_2 \oplus \cdots$.

The next question is how to use E and σ to understand Skl_3. Fortunately, one can create a noncommutative algebra from this data that is closely connected to Skl_3. This is the *twisted homogeneous coordinate ring* of E and is defined as follows. Let X be a k-scheme, with a line bundle $\mathcal{L}$ and automorphism σ. Set $\mathcal{L}_n = \mathcal{L} \otimes \mathcal{L}^{\sigma} \otimes \cdots \otimes \mathcal{L}^{\sigma^{n-1}}$, where $\mathcal{L}^{\tau} = \tau^* \mathcal{L}$ denotes the pull-back of $\mathcal{L}$ along an automorphism τ. Then the *twisted homogeneous coordinate ring* is defined to be the graded vector space $B = B(X, \mathcal{L}, \sigma) = k + \bigoplus_{n \geq 1} B_n$ where $B_n = \mathrm{H}^0(X, \mathcal{L}_n)$. The multiplication on $B = B(Y, \mathcal{L}, \sigma)$ is defined by the natural map

$$\begin{aligned} B_n \otimes_k B_m &\cong \mathrm{H}^0(X, \mathcal{L}_n) \otimes_k \sigma^n \mathrm{H}^0(X, \mathcal{L}_m) \\ &\cong \mathrm{H}^0(X, \mathcal{L}_n) \otimes_k \mathrm{H}^0(X, \mathcal{L}_m^{\sigma^n}) \xrightarrow{\phi} \mathrm{H}^0(X, \mathcal{L}_{n+m}) = B_{n+m}. \end{aligned}$$

The ring B has two significant properties. First, the way it has been constructed ensures that the natural isomorphism $S_1 \cong \mathrm{H}^0(\mathbb{P}^2, \mathcal{O}_{\mathbb{P}^2}(1)) \cong B_1$ induces a ring homomorphism $\phi : S \to B$. With a little more work using the Riemann-Roch theorem one can even show that $B \cong S/gS$ for some $g \in S_3$. Secondly—and this will be explained in more detail in the next section—$\mathrm{qgr}\, B \cong \mathrm{coh}(E)$. The latter fact allows one to obtain a detailed understanding of the structure of B and the former allows one to pull this information back to S.

To summarize, the point modules over the Sklyanin algebra Skl_3 are determined by an automorphism of an elliptic curve E and the geometry of E allows one to determine the structure of Skl_3. As is shown in [6] this technique works more generally and this leads to the following theorem.

Theorem 3 [6, 26, 27] *The AS-regular rings R of dimension 3 are classified. They are all noetherian domains with the Hilbert series of a weighted polynomial ring $k[x, y, z]$; thus the (x, y, z) can be given degrees (a, b, c) other than $(1, 1, 1)$.*

Moreover, R always maps homomorphically onto a twisted homogeneous coordinate ring $B = B(X, \mathcal{L}, \sigma)$, for some scheme X. Thus $\mathrm{coh}(X) \simeq \mathrm{qgr}\, B \hookrightarrow \mathrm{qgr}\, R$.

In this result, Artin, Tate and Van den Bergh [6] classified the algebras generated in degree one, while Stephenson [26, 27] did the general case.

There are strong arguments (see [11] or [25, Section 11]) for saying that the noncommutative analogues of the projective plane are precisely the categories $\mathrm{qgr}\, R$, where R is an AS-regular ring with the Hilbert series $1/(1-t)^3$ of the unweighted polynomial ring $k[x, y, z]$. So consider this class, which clearly includes the Sklyanin algebra. The second paragraph of the theorem can now be refined to say that *either* $X = \mathbb{P}^2$, in which case $\mathrm{qgr}\, R \simeq \mathrm{coh}(\mathbb{P}^2)$, *or* $X = E$ is a cubic curve in $\mathbb{P}^2$. Thus, the theorem can be interpreted as saying that *noncommutative projective planes are either equal to $\mathbb{P}^2$ or contain a commutative curve E.*

3. Twisted homogeneous coordinate rings

The ideas from [6] outlined in the last section have had many other applications, but before we discuss them we need to analyse twisted homogeneous coordinate rings in more detail. The following exercise may give the reader a feel for the construction.

Exercise 4 Perhaps the simplest algebra appearing in the theory of quantum groups is the quantum (affine) plane $k_q[x,y] = k\{x,y\}/(xy - qyx)$, for $q \in k^*$. Prove that $k_q[x,y] \cong B(\mathbb{P}^1, \mathcal{O}_{\mathbb{P}^1}(1), \sigma)$ where σ is defined by $\sigma(a:b) = (a:qb)$, for $(a:b) \in \mathbb{P}^1$.

For the rest of the section, fix a k-scheme X with an invertible sheaf $\mathcal{L}$ and automorphism σ. When $\sigma = 1$, the homogeneous coordinate ring $B(X,\mathcal{L}) = B(X,\mathcal{L},1)$ is a standard construction and one has Serre's fundamental theorem: If $\mathcal{L}$ is ample then $\mathrm{coh}(X) \simeq \mathrm{qgr}(B)$. As was hinted in the last section, this does generalize to the noncommutative case, provided one changes the definition of ampleness. Define $\mathcal{L}$ to be *σ-ample* if, for all $\mathcal{F} \in \mathrm{coh}(X)$, one has $\mathrm{H}^q(X, \mathcal{F} \otimes \mathcal{L}_n) = 0$ for all $q > 0$ and all $n \gg 0$. The naïve generalization of Serre's Theorem then holds.

Theorem 5 (Artin-Van den Bergh [7]) *Let X be a projective scheme with an automorphism σ and let $\mathcal{L}$ be a σ-ample invertible sheaf. Then $B = B(X,\mathcal{L},\sigma)$ is a right noetherian cg ring such that* $\mathrm{qgr}(B) \simeq \mathrm{coh}(X)$.

This begs the question of precisely which line bundles are σ-ample. A simple application of the Riemann-Roch Theorem shows that

$$\textit{if } X \textit{ is a curve, then any ample invertible sheaf is } \sigma\textit{-ample,} \tag{3.1}$$

and the converse holds for irreducible curves. This explains why Theorem 5 could be applied to the factor of the Sklyanin algebra in the last section.

For higher dimensional varieties the situation is more subtle and is described by the following result, for which we need some notation. Let X be a projective scheme and write $A^1_{\mathrm{Num}}(X)$ for the set of Cartier divisors of X modulo numerical equivalence. Let σ be an automorphism of X and let P_σ denote its induced action on $A^1_{\mathrm{Num}}(X)$. Since $A^1_{\mathrm{Num}}(X)$ is a finitely generated free abelian group, P_σ may be represented by a matrix and P_σ is called *quasi-unipotent* if all the eigenvalues of this matrix are roots of unity.

Theorem 6 (Keeler [15]) *If σ be an automorphism of a projective scheme X then:*

(1) X has a σ-ample line bundle if and only if P_σ is quasi-unipotent. If P_σ is quasi-unipotent, then all ample line bundles are σ-ample.

(2) In Theorem 5, B is also left noetherian.

There are two comments that should be made about Theorem 6. First, it is standard that $\mathrm{GKdim}\, B(X,\mathcal{L}) = 1 + \dim X$, whenever $\mathcal{L}$ is ample. However, it can happen that $\mathrm{GKdim}\, B(X,\mathcal{L},\sigma) > 1 + \dim X$. Secondly, one can still construct $B(X,\mathcal{L},\sigma)$ when $\mathcal{L}$ is ample but P_σ is not quasi-unipotent, but the resulting algebra is rather unpleasant. Indeed, possibly after replacing $\mathcal{L}$ by some $\mathcal{L}^{\otimes n}$, $B(X,\mathcal{L},\sigma)$ will be a non-noetherian algebra of exponential growth. See [15] for the details.

4. Noncommutative curves and surfaces

As we have seen, twisted homogeneous coordinate rings are fundamental to the study of noncommutative projective planes. However, a more natural starting place would be cg algebras of Gelfand-Kirillov dimension two since, as we suggested in the introduction, these should correspond to noncommutative curves. Their structure is particularly simple.

Theorem 7 [4] *Let R be a cg domain of GK-dimension 2 generated in degree one. Then there exists an irreducible curve Y with automorphism σ and ample invertible sheaf $\mathcal{L}$ such that R embeds into the twisted homogeneous coordinate ring $B(Y, \mathcal{L}, \sigma)$ with finite index. Equivalently, $R_n \cong \mathrm{H}^0(Y, \mathcal{L} \otimes \mathcal{L}^{\sigma} \otimes \cdots \otimes \mathcal{L}^{\sigma^{n-1}})$ for $n \gg 0$.*

By (3.1) we may apply Theorem 5 to obtain part (1) of the next result.

Corollary 8 *Let R be as in Theorem 7. Then:*

(1) R is a noetherian domain with $\operatorname{qgr} R \simeq \operatorname{coh}(Y)$. In particular, $\operatorname{qgr} R \simeq \operatorname{qgr} C$ for the commutative ring $C = B(Y, \mathcal{L}, \mathrm{Id})$.

(2) If $|\sigma| < \infty$ then R is a finite module over its centre. If $|\sigma| = \infty$, then R is a primitive ring with at most two height one prime ideals.

If R is not generated in degree one, then the analogue of Theorem 7 is more subtle, since more complicated algebras appear. See [4] for the details. One should really make a further generalization by allowing R to be prime rather than a domain and to allowing k to be arbitrary (since this allows one to consider the projective analogues of classical orders over Dedekind domains). Theorem 7 and Corollary 8 do generalize appropriately but the results are more technical. The details can be found in [5].

Although these results are satisfying they are really only half of the story. As in the commutative case one would also like to define noncommutative curves abstractly and then show that they can indeed be described by graded rings of the appropriate form. Such a result appears in [19] but to state it we need a definition.

Let $\mathcal{C}$ be an Ext-finite abelian category of finite homological dimension with derived category of bounded complexes $D^b(\mathcal{C})$. Recall that a cohomological functor $H : D^b(\mathcal{C}) \to \operatorname{mod}(k)$ is of *finite type* if, for $A \in D^b(\mathcal{C})$, only a finite number of the $H(A[n])$ are non-zero. The category $\mathcal{C}$ is *saturated* if every cohomological functor $H : D^b(\mathcal{C}) \to \operatorname{mod}(k)$ of finite type is of the form $\operatorname{Hom}(A, -)$ (that is, H is representable). If X is a smooth projective scheme, then $\operatorname{coh}(X)$ is saturated [10], so it is not unreasonable to use this as part of the definition of a "noncommutative smooth curve."

Theorem 9 (Reiten-Van den Bergh [19, Theorem V.1.2]) *Assume that $\mathcal{C}$ is a connected saturated hereditary noetherian category. Then $\mathcal{C}$ has one of the following forms:*

(1) $\operatorname{mod}(\Lambda)$ where Λ is an indecomposable finite dimensional hereditary algebra.

(2) $\operatorname{coh}(\mathcal{O})$ where $\mathcal{O}$ is a sheaf of hereditary $\mathcal{O}_X$-orders over a smooth connected projective curve X.

It is easy to show that the abelian categories appearing in parts (1) and (2) of this theorem are of the form $\operatorname{qgr} R$ for a graded ring R with $\operatorname{GKdim} R \leq 2$, and so this result can be regarded as a partial converse to Theorem 7. A discussion of the saturation condition for noncommutative algebras may be found in [12].

If one accepts that noncommutative projective curves and planes have been classified, as we have argued, then the natural next step is to attempt to classify all noncommutative surfaces and this has been a major focus of recent research. This program is discussed in detail in [25, Sections 8–13] and so here we will be very brief. For the sake of argument we will assume that an (irreducible) noncommutative surface is $\operatorname{qgr} R$ for a noetherian cg domain R with $\operatorname{GKdim} R = 3$, although the precise definition is as yet unclear. For example, Artin [1] demands that $\operatorname{qgr} R$ should also possess a dualizing complex in the sense of Yekutieli [30]. Nevertheless in attempting to classify surfaces it is natural to mimic the commutative proof:

(a) Classify noncommutative surfaces up to birational equivalence; equivalently classify the associated graded division rings of fractions for graded domains R with $\operatorname{GKdim} R = 3$. Artin [1, Conjecture 4.1] conjectures that these division rings are known.
(b) Prove a version of Zariski's theorem that asserts that one can pass from any smooth surface to a birationally equivalent one by successive blowing up and down. Then find minimal models within each equivalence class.

Van den Bergh has created a noncommutative theory of blowing up and down [28, 29] and used this to answer part (b) in a number of special cases. A key fact in his approach is that (after minor modifications) each known example of a noncommutative surface $\operatorname{qgr} R$ contains an embedded commutative curve $\mathcal{C}$, just as $\operatorname{qgr}(\mathrm{Skl}_3) \hookleftarrow \operatorname{coh}(E) = E$ in Section 2. This is important since he needs to blow up points on that subcategory. In general, define *a point in* $\operatorname{qgr} R$ to be $\pi(M)$ for a point module $M \in \operatorname{gr} R$. Given such a point p, write $p = \pi(R/I) = \mathcal{R}/\mathcal{I}$. Mimicking the classical situation we would like to write

$$\mathcal{B} = \mathcal{R} \oplus \mathcal{I} \oplus \mathcal{I}^2 \oplus \cdots, \tag{4.1}$$

and then define the blow-up of $\operatorname{qgr} R$ to be the category $\operatorname{qgr} \mathcal{B}$ of finitely generated graded $\mathcal{B}$-modules modulo those that are right bounded. However, there are two problems. A minor one is that $\mathcal{I}$ needs to be twisted to take into account the shift functor on $\operatorname{qgr} R$. The major one is that I is only a one-sided ideal of R, and so there is no natural multiplication on $\mathcal{B}$. To circumvent these problems, Van den Bergh [28] has to define $\mathcal{B}$ in a more subtle category so that it is indeed an algebra. It is then quite hard to prove that $\operatorname{qgr} \mathcal{B}$ has the appropriate properties.

5. Hilbert schemes

Since point modules and twisted homogeneous coordinate rings have proved so useful, it is natural to ask how generally these techniques can be applied. In particular, one needs to understand when point modules, or other classes of modules, can be parametrized by a scheme. Indeed, even for point modules over surfaces the

answer was unknown until recently and this is obviously rather important for the program outlined in the last section.

The best positive result is due to Artin, Small and Zhang [3, 9], for which we need a definition. A k-algebra R is called *strongly noetherian* if $R \otimes_k C$ is noetherian for all noetherian commutative k-algebras C.

Theorem 10 (Artin-Zhang [9, Theorems E4.3 and E4.4]) *Assume that R is a strongly noetherian, cg algebra and fix $h(t) = \sum h_i t^i \in k[[t]]$. Let $\mathcal{C}$ denote the set of cyclic R-modules $M = R/I$ with Hilbert series $h_M(t) = \sum \dim_k(M_i)t^i$ equal to $h(t)$. Then:*

(1) $\mathcal{C}$ is naturally parametrized by a (commutative) projective scheme.
(2) There exists an integer d such that, if $M = R/I \in \mathcal{C}$, then I is generated in degrees $\leq d$ as a right ideal of R.

In particular, if R is a strongly noetherian cg algebra generated in degree one, then the set of point modules is naturally parametrized by a projective scheme $\mathcal{P}$. In this case one can further show that *the shift functor $M \mapsto M_{\geq 1}[1]$ induces an automorphism σ of $\mathcal{P}$*. Thus one can form the corresponding twisted homogeneous coordinate rings $B = B(\mathcal{P}, \mathcal{L}, \sigma)$ and for an appropriate line bundle $\mathcal{L}$ there will exist a homomorphism $\phi : R \to B$. Determining when ϕ is surjective is probably quite subtle. This result cannot be used to shorten the arguments about the Sklyanin algebra Skl_3 given in Section 2, since one needs to use $B(E, \mathcal{L}, \sigma)$ to prove that Skl_3 is noetherian.

Although we have concentrated on point modules, more general classes of modules are also important. An example where *line modules* (modules M with the Hilbert series of $k[x, y]$) are needed in a classification problem appears in [22].

How strong is the strongly noetherian hypothesis? Certainly most of the standard examples of noetherian cg algebras (including the Sklyanin algebras) are strongly noetherian (see [3, Section 4]) and so one might hope that this is always the case. But in fact, as Rogalski [20] has shown, cg noetherian algebras that are not strongly noetherian exist in profusion.

These examples are constructed as subrings of $B = B(\mathbb{P}^n, \mathcal{O}_{\mathbb{P}^n}(1), \sigma)$ for an appropriate automorphism σ. Given $\sigma \in \mathrm{Aut}(\mathbb{P}^n)$, pick $c \in \mathbb{P}^n$ and set $\mathcal{C} = \{c_i = \sigma^{-i}(c) : i \in \mathbb{N}\}$. Then $\mathcal{C}$ is called *critically dense* if, for any infinite subset $\mathcal{D} \subseteq \mathcal{C}$, the Zariski closure of $\mathcal{D}$ equals $\mathbb{P}^n$. This is not a particularly stringent condition, since it holds for a generic set of $(\sigma, c) \in \mathrm{Aut}(\mathbb{P}^n) \times \mathbb{P}^n$. Corresponding to c one has the point module $M = B/VB$ for some codimension one subspace $V = V(c) \subseteq B_1$. Rogalski's example is then simply $S(\sigma, c) = k\langle V \rangle \subset B$, and it has remarkable properties:

Theorem 11 (Rogalski [20]) *Keep the above notation. Assume that $\sigma \in \mathrm{Aut}(\mathbb{P}^n)$ and $c \in \mathbb{P}^n$ for $n \geq 2$ are such that $\mathcal{C}$ is critically dense. Then:*

(1) $S = S(\sigma, c)$ is always noetherian but never strongly noetherian.
(2) The point modules for S are not naturally parametrized by a projective scheme.
(3) S satisfies the condition χ_1 but not the condition χ_2, as defined below. Moreover, $\mathrm{qgr}\, S$ has finite cohomological dimension.

(4) The category $\operatorname{qgr} S$ *is not Ext-finite; indeed if* $\mathcal{S} = \pi(S) \in \operatorname{qgr} S$, *then* $\mathrm{H}^1(\mathcal{S}) = \operatorname{Ext}^1_{\operatorname{qgr} S}(\mathcal{S}, \mathcal{S})$ *is infinite dimensional.*

Some comments about the theorem are in order. First, the point modules for $S = S(\sigma, c)$ are actually parametrized by an "infinite blowup of $\mathbb{P}^n$" in the sense that they are parametrized by $\mathbb{P}^n$ except that for each $p \in \mathcal{C}$ one has a whole family $\mathcal{P}_p$ of point modules parametrized by $\mathbb{P}^{n-1}$. In contrast, the points in $\operatorname{qgr} S$ are actually parametrized by $\mathbb{P}^n$ since, if $M, N \in \mathcal{P}_p$, then $\pi(N) \cong \pi(M)$ in $\operatorname{qgr} S$.

The conditions χ_i in part (3) are defined as follows: A cg ring R satisfies χ_n if, for each $0 \leq j \leq n$ and each $M \in \operatorname{gr} R$, one has $\dim_k \operatorname{Ext}^j_R(k, M) < \infty$. The significance of χ_1 is that, by [8, Theorem 4.5], one can reconstruct $S = S(\sigma, c)$ from $\operatorname{qgr} S$ and so the peculiar properties of S are reflected in $\operatorname{qgr} S$. In particular, part (4) implies that S does not satisfy χ_2. The significance of part (4) is that, for all the algebras R considered until now, Serre's finiteness theorem holds in the sense that $\mathrm{H}^i(\mathcal{F})$ is finite dimensional for all $\mathcal{F} \in \operatorname{qgr} R$ and all i.

Here is the simplest example of $S(\sigma, c)$. Pick algebraically independent elements $p, q \in k$ and define $\sigma \in \operatorname{Aut}(\mathbb{P}^2)$ by $\sigma(a:b:c) = (pa:qb:c)$. If $c = (1:1:1) \in \mathbb{P}^2$ then $\mathcal{C}$ is critically dense and an argument like that of Exercise 4 shows that

$$B = k\{x, y, z\}/(zx - pxz,\ zy - qyz,\ yx - pq^{-1}xy) \quad \text{and} \quad S(\sigma, c) = k\langle y - x,\ z - x\rangle.$$

This example was first considered by Jordan [14] who was able to parametrize the point modules for $S(\sigma, c)$ but was unable to determine if the ring was noetherian.

Rogalski's examples show that, even for surfaces, the picture is much more complicated than the discussion of the last section would suggest. Yet even these examples appear in a geometric framework; indeed they can be constructed as blow-ups of $\mathbb{P}^n$ if one uses the naïve approach of (4.1).

This works as follows. As before, assume that $(\sigma, c) \in \operatorname{Aut}(\mathbb{P}^n) \times \mathbb{P}^n$ for $n \geq 2$ is such that $\mathcal{C}$ is critically dense. In $\operatorname{coh}(\mathbb{P}^n)$ let $\mathcal{I}_c$ denote the ideal sheaf corresponding to the point c. If $\mathcal{L}$ is a coherent module over $\mathcal{O} = \mathcal{O}_{\mathbb{P}^n}$, we form a bimodule $\mathcal{L}_\sigma$ such that as a left module, $\mathcal{L}_\sigma \cong \mathcal{L}$ but the right action is twisted by σ: if $s \in \mathcal{L}_\sigma(U)$ and $a \in \mathcal{O}_{\mathbb{P}^n}(\sigma U)$, then $sa \in \mathcal{L}_\sigma(U)$ is defined by the formula $sa = a^\sigma s$. See [7, pp.252-3] for a more formal discussion. Now set $\mathcal{J} = \mathcal{I}_c \otimes_{\mathcal{O}} \mathcal{O}(1)_\sigma \subseteq \mathcal{O}(1)_\sigma$ and let $\mathcal{B} = \mathcal{B}(\sigma, c) = \mathcal{O} \oplus \mathcal{J} \oplus \mathcal{J}^2 \oplus \cdots$, where $\mathcal{J}^n$ is the image of $\mathcal{J}^{\otimes n}$ in $\mathcal{O}(1)_\sigma^{\otimes n} \cong \mathcal{O}(n)_{\sigma^n}$. This does not define a sheaf of rings in the usual sense since we are "playing a game of musical chairs with the open sets [7, p.252]." Nevertheless $\mathcal{B}$ does have an natural graded algebra structure and so we can form $\operatorname{qgr} \mathcal{B}$ in the usual way. If $\sigma = 1$ then $\operatorname{qgr} \mathcal{B}$ is simply $coh(X)$, where X is the blow-up of $\mathbb{P}^n$ at c. In contrast, Keeler, Rogalski and the author have recently proved:

Theorem 12 [16] *Pick* $(\sigma, c) \in \operatorname{Aut}(\mathbb{P}^n) \times \mathbb{P}^n$ *for* $n \geq 2$ *such that* $\mathcal{C}$ *is critically dense. Then* $\mathcal{B} = \mathcal{B}(\sigma, c)$ *is noetherian. Moreover* $\operatorname{qgr}(\mathcal{B}) \simeq \operatorname{qgr} S(\sigma, c)$.

Thus, $\operatorname{qgr} S(\sigma, c)$ is nothing more than the (noncommutative) blow-up of $\mathbb{P}^n$ at a point! The differences between this blow-up and Van den Bergh's are illustrative. Van den Bergh had to work hard to ensure that the analogue of the exceptional divisor really looks like a curve. Indeed much of his formalism is required for just

this reason. In contrast, in Theorem 12 the analogue of the exceptional divisor (which in this case equals $\mathcal{B}/(\mathcal{I}_{c-1})\mathcal{B}$) is actually a point. This neatly explains the structure of the points in $\operatorname{qgr} S(\sigma, c)$; they are indeed parametrized by $\mathbb{P}^n$ although the point corresponding to c (and hence the shifts of this point, which are nothing more than the points corresponding to the c_i) are distinguished.

References

[1] M. Artin, Some problems on three-dimensional graded domains, *Representation theory and algebraic geometry*, London Math. Soc. Lecture Note Ser., vol. 238, Cambridge Univ. Press, Cambridge, 1995, 1–19.

[2] M. Artin and W. Schelter, Graded algebras of global dimension 3, *Adv. in Math.*, 66 (1987), 171–216.

[3] M. Artin, L. W. Small and J. J. Zhang, Generic flatness for strongly noetherian algebras. *J. Algebra*, 221 (1999), 579–610.

[4] M. Artin and J. T. Stafford, Noncommutative graded domains with quadratic growth, *Invent. Math.*, 122 (1995), 231–276.

[5] M. Artin and J. T. Stafford, Semiprime graded algebras of dimension two, *J. Algebra*, 277 (2000), 68–123.

[6] M. Artin, J. Tate, and M. Van den Bergh, Some algebras associated to automorphisms of elliptic curves, *The Grothendieck Festschrift*, vol. 1, Birkhäuser, Boston, 1990, 33–85.

[7] M. Artin and M. Van den Bergh, Twisted homogeneous coordinate rings, *J. Algebra*, 133 (1990), 249–271.

[8] M. Artin and J. J. Zhang, Noncommutative projective schemes, *Adv. in Math.*, 109 (1994), 228–287.

[9] M. Artin and J. J. Zhang, Abstract Hilbert schemes, *Algebr. Represent. Theory*, 4 (2001), 305–394.

[10] A. I. Bondal and M. M. Kapranov, Representable functors, Serre functors, and reconstructions, *Math. USSR-Izv.*, 35 (1990), 519–541.

[11] A. I. Bondal and A. E. Polishchuk, Homological properties of associative algebras: the method of helices, *Russian Acad. Sci. Izv. Math.*, **42** (1994), 219–260.

[12] A. I. Bondal and M. Van den Bergh, Generators and representability of functors in commutative and noncommutative geometry; math.AG/0204218 (to appear).

[13] A. Connes and M. Dubois-Violette, Noncommutative finite-dimensional manifolds. I. Spherical manifolds and related examples; math.QA/0107070 (to appear).

[14] D. A. Jordan, The graded algebra generated by two Eulerian derivatives, *Algebr. Represent. Theory*, 4 (2001), 249–275.

[15] D. S. Keeler, Criteria for σ-ampleness, *J. Amer. Math. Soc.*, 13 (2000), 517–532.

[16] D. S. Keeler, D. Rogalski and J. T. Stafford, work in progress.

[17] G. R. Krause and T. H. Lenagan, *Growth of algebras and Gelfand-Kirillov dimension*, Research Notes in Mathematics, vol. 116, Pitman, Boston, 1985.

[18] A. V. Odesskii and B. L. Feigin, Sklyanin's elliptic algebras, *Functional Anal. Appl.*, 23 (1989), no. 3, 207–214.
[19] I. Reiten and M. Van den Bergh, Noetherian hereditary categories satisfying Serre duality, *J. Amer. Math. Soc.*, 15 (2002), 295–366.
[20] D. Rogalski, Examples of generic noncommutative surfaces; math.RA/0203180 (to appear).
[21] A. L. Rosenberg, The spectrum of abelian categories and reconstruction of schemes, *Rings, Hopf algebras, and Brauer groups*, Lecture Notes in Pure and Appl. Math., vol. 197, Marcel Dekker, New York, 1998, 257–274.
[22] B. Shelton and M. Vancliff, Schemes of line modules I, *J. London Math. Soc.*; www.uta.edu/math/vancliff/R/ (to appear).
[23] E. K. Sklyanin, Some algebraic structures connected to the Yang-Baxter equation, *Functional Anal. Appl.*, 16 (1982), 27–34.
[24] J. T. Stafford and J. J. Zhang, Examples in noncommutative projective geometry, *Math. Proc. Cambridge Philos. Soc.*, 116 (1994), 415–433.
[25] J. T. Stafford and M. Van den Bergh, Noncommutative curves and noncommutative surfaces, *Bull. Amer. Math. Soc.*, 38 (2001), 171–216.
[26] D. R. Stephenson, Artin-Schelter regular algebras of global dimension three, *J. Algebra*, 183 (1996), 55–73.
[27] D. R. Stephenson, Algebras associated to elliptic curves, *Trans. Amer. Math. Soc.*, 349 (1997), 2317–2340.
[28] M. Van den Bergh, Blowing up of noncommutative smooth surfaces, *Mem. Amer. Math. Soc.*, 154 (2001), no. 734.
[29] M. Van den Bergh, Abstract blowing down, *Proc. Amer. Math. Soc.*, 128 (2000), 375–381.
[30] A. Yekutieli, Dualizing complexes over noncommutative graded algebras, *J. Algebra*, 153 (1992), 41–84.

ICM 2002 · Vol. II · 105–116

Deformations of Chiral Algebras

Dimitri Tamarkin*

Abstract

We start studying chiral algebras (as defined by A. Beilinson and V. Drinfeld) from the point of view of deformation theory. First, we define the notion of deformation of a chiral algebra on a smooth curve X over a bundle of local artinian commutative algebras on X equipped with a flat connection (whereas 'usual' algebraic structures are deformed over a local artinian algebra) and we show that such deformations are controlled by a certain *-Lie algebra $\mathfrak{g}$. Then we try to contemplate a possible additional structure on $\mathfrak{g}$ and we conjecture that this structure up to homotopy is a chiral analogue of Gerstenhaber algebra, i.e. a coisson algebra with odd coisson bracket (in the terminology of Beilinson-Drinfeld). Finally, we discuss possible applications of this structure to the problem of quantization of coisson algebras.

2000 Mathematics Subject Classification: 14, 18.

1. Introduction

Chiral algebras were introduced in [1]. In the same paper the authors introduced the classical limit of a chiral algebra which they call a coisson algebra and posed the problem of quantization of coisson algebras. The goal of this paper is to show how the theory of deformation quantization (=the theory of deformations of associative algebras of a certain type) in the spirit of [3] can be developed in this situation.

Central object in the theory of deformations of associative algebras is the differential graded Lie algebra of Hochschild cochains. It turns out that in our situation it is more appropriate to use what we call pro-*-Lie-algebras rather than usual Lie algebras (the notion of *-Lie algebra was also introduced in [1]). Next, we compute the cohomology of the pro-*-Lie-algebra controlling chiral deformations of a free commutative $\mathcal{D}_X$-algebra SK, where K is a locally free $\mathcal{D}_X$-module.

Next, we state an analogue of Gerstenhaber theorem which says that the cohomology of the deformation complex of an associative algebra carries the structure of a Gerstenhaber algebra. We give a definition of a chiral analogue of Gerstenhaber

*Department of Mathematics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA. E-mail: tamarkin@math.harvard.edu

algebra and define the operations of this structure on deformation pro-*-Lie algebra of a chiral algebra.

Finally, mimicking Kontsevich's formality theorem, we formulate the formality conjecture for the deformation pro-*-Lie algebra of the chiral algebra SK mentioned above and claim that this conjecture implies a 1-1 correspondence between deformations of SK and coisson brackets on SK.

2. Chiral algebras and their deformations

2.1. Chiral operations

In [1] chiral operations are defined as follows. Let X be a smooth curve and M_i,N $\mathcal{D}_X$-modules. Denote by $i_n : X \to X^n$ the diagonal embedding and by $j_n : U_n \to X^n$ the open embedding of the complement to all diagonals in X^n. Set

$$P_{\mathbf{ch}}(M_1, \dots, M_n; N) = \hom_{\mathcal{D}_{X^n}}(j_* j^*(M_1 \boxtimes \cdots \boxtimes M_n), i_{n*} N). \tag{1}$$

In the case $n = 0$ set

$$P_{\mathbf{ch}}(M) = H^0(M \otimes_{\mathcal{D}_X} \mathcal{O}_X).$$

Let M be a fixed D_X-module. Write

$$P_{\mathbf{ch}M}(n) = P_{\mathbf{ch}}(M, M, \dots, M; M).$$

It is explained in [1] that P_M is an operad.

2.1.1. Chiral algebras

Let **lie** be the operad of Lie algebras. *A chiral algebra* structure on M is a homomorphism $\mathbf{lie} \to P_M$. We have a standard chiral algebra structure on $M = \omega_X$. A chiral algebra M is called unital if it is endowed with an injection $\omega_X \to M$ of chiral algebras.

2.2. Deformations

2.2.1. Agreements

To simplify the exposition, we will only consider unital chiral algebras M with the following restrictions: we assume that X is affine and the D_X-module M can be represented as $M = \omega_X \oplus N$, where $N \cong E \otimes_{\mathcal{O}_X} D_X$ for some locally free coherent sheaf E.

2.2.2. Nilpotent $\mathcal{D}_X$-algebras

Let E be a left $\mathcal{D}_X$-module equipped with a commutative associative unital product $E \otimes E \to E$. Let $u : \mathcal{O}_X \to E$ be the unit embedding. Call E nilpotent if there exists a $\mathcal{D}_X$-module splitting $s : E \cong \mathcal{M} \oplus \mathcal{O}_X$ and a positive integer N such that the N-fold product vanishes on $\mathcal{M}$. $\mathcal{M}$ is then a unique maximal $\mathcal{D}_X$-ideal in E.

2.2.3. Deformations over a nilpotent $\mathcal{D}_X$-algebra

Let E be a nilpotent $\mathcal{D}_X$-algebra with maximal ideal $\mathcal{M}$. We have a notion of E-module and of an E-linear chiral algebra . For any $\mathcal{D}_X$-module M, $M_E := M \otimes_{\mathcal{O}_X} E$ is an E-module.

Let M be a chiral algebra. An E-linear unital chiral algebra structure on M_E is called *deformation of M over E* if the induced structure on $M_E/\mathcal{M}.M_E \cong M$ coincides with the one on M. Denote by $G_M(E)$ the set of all isomorphism classes of such deformations.

2.3. The functor G_M and its representability

It is clear that $E \mapsto G_M(E)$ is a functor from the category of nilpotent D_X-algebras to the category of sets. In classical deformation theory one usually has a functor from the category of (usual) local Arminian (=nilpotent and finitely dimensional) algebras to the category of sets and one tries to represent it by a differential graded Lie algebra. In this section we will see that in our situation a natural substitute for a Lie algebra is a so-called *-Lie algebra in the sense of [1]. More precise, given a *-Lie algebra $\mathfrak{g}$, we are going to construct a functor $F_{\mathfrak{g}}$ from the category of nilpotent $\mathcal{D}_X$-algebras to the category of sets. In the next section we will show that the functor G_M is 'pro-representable' in this sense. We will construct a pro-*-Lie algebra $\mathbf{def}_M$ (exact meaning will be given below) and an isomorphism of functors G_M and $F_{\mathbf{def}_M}$.

2.3.1. *-Lie algebras

[1] Let $\mathfrak{g}_i$, N be right $\mathcal{D}_X$-modules. Set

$$P_*(\mathfrak{g}_1, \dots, \mathfrak{g}_n; N) := \hom_{\mathcal{D}_{X^n}}(\mathfrak{g}_1 \boxtimes \cdots \boxtimes \mathfrak{g}_n, i_{n*}N),$$

and $P_{*\mathfrak{g}}(n) := P(\mathfrak{g}, \dots, \mathfrak{g}; \mathfrak{g})$. It is known that $P_{*\mathfrak{g}}$ is an operad. A *-Lie algebra structure on $\mathfrak{g}$ is by definition a morphism of operads $f : \mathbf{lie} \to P_{*\mathfrak{g}}$. Let $b \in \mathbf{lie}(2)$ be the element corresponding to the Lie bracket. We call $f(b) \in P_{*\mathfrak{g}}(2)$ the *-Lie bracket.

2.3.2.

Let $\mathfrak{g}$ be a *-Lie algebra and A be a commutative $\mathcal{D}_X$-algebra. introduce a vector space $\mathfrak{g}(A) = \mathfrak{g} \otimes_{\mathcal{D}_X} A$. This space is naturally a Lie algebra. Indeed, we have a *-Lie bracket $\mathfrak{g} \boxtimes \mathfrak{g} \to i_{2*}\mathfrak{g}$. Multiply both parts by $A \boxtimes A$:

$$(\mathfrak{g} \boxtimes \mathfrak{g}) \otimes_{\mathcal{D}_{X\times X}} (A \boxtimes A) \to i_{2*}\mathfrak{g} \otimes_{\mathcal{D}_{X\times X}} (A \boxtimes A). \qquad (*)$$

The left hand side is isomorphic to $\mathfrak{g}(A) \otimes \mathfrak{g}(A)$. The right hand side is isomorphic to $\mathfrak{g} \otimes_{\mathcal{D}_X} (A \otimes_{\mathcal{O}_X} A)$. Thus, (*) becomes:

$$\mathfrak{g}(A) \otimes \mathfrak{g}(A) \to \mathfrak{g} \otimes_{\mathcal{D}_X} (A \otimes_{\mathcal{O}_X} A).$$

The product on A gives rise to a map

$$\mathfrak{g} \otimes_{\mathcal{D}_X} (A \otimes_{\mathcal{O}_X} A) \to \mathfrak{g} \otimes_{\mathcal{D}_X} A \cong \mathfrak{g}(A),$$

and we have a map $\mathfrak{g}(A) \otimes \mathfrak{g}(A) \to \mathfrak{g}(A)$. It is straightforward to check that this map is a Lie bracket.

2.3.3.

Let now $\mathfrak{g}$ be a differential graded *-Lie algebra and let A be a differential graded commutative $\mathcal{D}_X$-algebra. Then $\mathfrak{g}(A) := \mathfrak{g} \otimes_{\mathcal{D}_X} A$ is a differential graded Lie algebra.

2.3.4.

Let A be a nilpotent $\mathcal{D}_X$ algebra and $\mathcal{M}_A \subset A$ be the maximal nilpotent ideal. Then $\mathfrak{g}(\mathcal{M}_A)$ is a nilpotent differential graded Lie algebra.

2.3.5.

Recall that given a differential graded nilpotent Lie algebra $\mathfrak{n}$, one can construct the so called Deligne groupoid $\mathcal{G}_\mathfrak{n}$. Its objects are all $x \in \mathfrak{n}^1$ satisfying $dx + [x,x]/2 = 0$ (so called Maurer-Cartan elements). The group $\exp(\mathfrak{n}^0)$ acts on the set of Maurer-Cartan elements by gauge transformations. $\mathcal{G}_\mathfrak{n}$ is the groupoid of this action. Denote by $\mathcal{D}_\mathfrak{n}$ the set of isomorphism classes of this groupoid. If $f : \mathfrak{n} \to \mathfrak{m}$ is a map of differential graded Lie algebras such that the induced map on cohomology $H^i(f)$ is an isomorphism for all $i \geq 0$, then the induced map $\mathcal{D}_\mathfrak{n} \to \mathcal{D}_\mathfrak{m}$ is a bijection. If $\mathfrak{n}, \mathfrak{m}$ are both centered in non-negative degrees, then the induced map $\mathcal{G}_\mathfrak{n} \to \mathcal{G}_\mathfrak{m}$ is an equivalence of categories. Since in our situation we will deal with Lie algebras centered in arbitrary degrees, we will use $\mathcal{D}_\mathfrak{n}$ rather than groupoids.

2.3.6.

Set $F_M(A) = \mathcal{D}_{\mathfrak{g}(\mathcal{M}_A)}$. It is a functor from the category of nilpotent $\mathcal{D}_X$-algebras to the category of sets.

2.4. Pro-*-Lie- algebras

*-Lie algebras are insufficient for description of deformations of chiral algebras. We will thus develop a generalization. We need some preparation

2.4.1. Procategory

For an Abelian category C consider the category **pro** C whose objects are functors $I \to C$, where I is a small filtered category. Let $F_k : I_k \to C$, $k = 1, 2$ be objects. Set

$$\hom(F_1, F_2) := \mathrm{liminv}_{i_2 \in I_2} \mathrm{limdir}_{i_1 \in I_1} (F_1(i_1), F_2(i_2)).$$

The composition of morphisms is naturally defined. One can show that **pro** C is an Abelian category. Objects of **pro** C are called pro-objects.

2.4.2. Direct image of pro-$\mathcal{D}$-modules

Let $M : I \to \mathcal{D}_Y - mod$ be a pro-object, where Y is a smooth algebraic variety and let $f : Y \to Z$ be a locally closed embedding. Denote the composition $f_* \circ M : I \to \mathcal{D}_Z - mod$ simply by f_*M. We will get a functor $f_* : \mathbf{pro}\ \mathcal{D}_Y - mod \to \mathbf{pro}\ \mathcal{D}_Z - mod$.

2.4.3. Chiral and *-operations

For $N, M_i \in \mathbf{pro}\ \mathcal{D}_X - mod$ we define $P_*(M_1, \ldots, M_n, N)$, $P_{\mathbf{ch}}(M_1, \ldots, M_n, N)$ by exactly the same formulas as for usual $\mathcal{D}_X$-modules.

2.4.4. pro-*-Lie algebras

*-Lie algebra structure on a pro-$\mathcal{D}_X$-module is defined in the same way as for usual $\mathcal{D}_X$-modules.

2.4.5.

For a pro-right $\mathcal{D}_X$-module $I \to M$ and a left $\mathcal{D}_X$-module L define a vector space $M \otimes_{\mathcal{D}_X} L = \mathrm{liminv}_I(M \otimes_{\mathcal{D}_X} L)$. For a *-Lie algebra $\mathfrak{g}$ and a commutative $\mathcal{D}_X$-algebra a, $\mathfrak{g} \otimes_{\mathcal{D}_X} a$ is a Lie algebra. Construction is the same as for usual *-Lie algebras. Similarly, we can define the functor $F_{\mathfrak{g}}$ from the category of nilpotent $\mathcal{D}_X$-algebras to the category of sets.

2.5. Representability of G_M by a pro-*-Lie algebra

We are going to construct a differential graded *-pro-Lie algebra $\mathfrak{g}$ such that $F_{\mathfrak{g}}$ is equivalent to $G_{\mathcal{M}}$. We need a couple of technical lemmas.

2.5.1.

Let Y be a smooth affine algebraic varieties and U, V be right $\mathcal{D}_Y$-modules. Let $U_\alpha, \alpha \in A$ be the family of all finitely generated submodules of U. Denote $\mathbf{prohom}(U, V) = \mathrm{liminv}_\alpha(U_\alpha, V)$ the corresponding pro-vector space.

2.5.2.

Let $i : X \to Y$ be a closed embedding, let B be a $\mathcal{D}_Y$-module and M be a $\mathcal{D}_X$-module. Then

$$\mathbf{prohom}_{\mathcal{D}_Y}(B, i_*(M \otimes_{\mathcal{O}_X} \mathcal{D}_X))$$

is a pro-$\mathcal{D}_X$-module. Denote it by $P(B, M)$. Let now $Y = X^n$.

Lemma 2.1 *Assume that $B = j_{n*}j_n^*(E \otimes_{\mathcal{O}_{X^n}} \mathcal{D}_{X^n})$, where E is locally free and coherent. For any left $\mathcal{D}_X$-module L we have*

$$\mathbf{prohom}(B, i_{n*}(M \otimes_{\mathcal{O}_X} L)) \cong P(B, M) \otimes_{\mathcal{D}_X} L.$$

Proof. Let $F = j_{n*}j_n^*E$. We have $B = F \otimes_{\mathcal{O}_{X^n}} \mathcal{D}_{X^n}$. Note that $F = \text{limdir } F_\alpha$, where F_α runs through the set of all free coherent submodules of F.

We have

$$\begin{aligned} P(B,M) &= \text{liminv } \hom_{\mathcal{D}_{X^n}}(F_\alpha \otimes_{\mathcal{O}_{X^n}} \mathcal{D}_{X^n}, i_{n*}(M \otimes_{\mathcal{O}_X} L)) \\ &\cong \text{liminv } F_\alpha^* \otimes_{\mathcal{O}_{X^n}} i_{n*}(M \otimes_{\mathcal{O}_X} \mathcal{D}_X) \otimes_{\mathcal{D}_X} L \\ &\cong \text{liminv } \hom_{\mathcal{O}_{X^n}}(F_\alpha, i_{n*}(M \otimes_{\mathcal{O}_X} \mathcal{D}_X)) \otimes_{\mathcal{D}_X} L \\ &\cong \mathbf{prohom}(B, i_{n*}(M \otimes_{\mathcal{O}_X} \mathcal{D}_X)) \otimes_{\mathcal{D}_X} L. \end{aligned}$$

2.5.3.

Let B, M be as above. We have a natural morphism

$$i : i_{n*}P(B,M) \cong P(B,M) \otimes_{\mathcal{D}_X} \mathcal{D}_X^{\otimes^n_{\mathcal{O}_X}} \to \mathbf{prohom}(B, M \otimes_{\mathcal{D}_X} \otimes (\mathcal{D}_X)^{\otimes^n_{\mathcal{O}_X}}).$$

The above lemmas imply that i is an isomorphism.

2.5.4.

Let M be a right $\mathcal{D}_X$-module. Set

$$U_M(n) = \mathcal{P}_{\mathbf{ch}}(M, M, \dots, M; M \otimes \mathcal{D}_X) := \mathbf{prohom}(j_{n*}j_n^* M^{\boxtimes n}, i_{n*}(M \otimes \mathcal{D}_X)),$$

it is a right pro-$\mathcal{D}_X$-module. We will endow the collection U_M with the structure of an operad in *-pseudotensor category. This means that we will define the composition maps

$$\circ_i \in P_*(U_M(n), U_M(m); U_M(n+m-1)),$$

$i = 1, \dots, n+m-1$, satisfying the operadic axioms. We need a couple of technical facts.

2.5.5.

Let $i_n : X \to X^n$ be the diagonal embedding and $p_n^i : X^n \to X$ be the projections. Lemma 2.5.3. implies that

Lemma 2.2

$$i_{n*}U_M(k) \cong \mathcal{P}_{\mathbf{ch}}(M, \dots, M; M \otimes_{\mathcal{D}_X} \mathcal{D}_X^{\otimes n}).$$

Lemma 2.3 *For any $\mathcal{D}_X$-modules M, S we have an isomorphism*

$$i_{n*}(M) \otimes p_n^{i*}S \cong i_{n*}(M \otimes S).$$

2.5.6.

We are now ready to define the desired structure. In virtue of 2.3 we have natural maps:

$$\mathcal{P}_{\mathbf{ch}}(M_1,\dots,M_n;N)\to\mathcal{P}_{\mathbf{ch}}(M_1,\dots,M_i\otimes\mathcal{D}_X,\dots,M_n;(N\otimes\mathcal{D}_X)).$$

Thus, we have maps:

$$\begin{aligned}
&\mathcal{P}_{\mathbf{ch}}(M_1,\dots,M_n;(N_i\otimes\mathcal{D}_X))\boxtimes\mathcal{P}_{\mathbf{ch}}(N_1,\dots N_m;(K\otimes\mathcal{D}_X))\\
\to\ &\mathcal{P}_{\mathbf{ch}}(M_1,\dots,M_n;(N_i\otimes\mathcal{D}_X))\\
&\boxtimes\mathcal{P}_{\mathbf{ch}}(N_1,\dots,N_i\otimes\mathcal{D}_X,\dots,N_m;(K\otimes\mathcal{D}_X)\otimes\mathcal{D}_X)\\
\to\ &\mathcal{P}_{\mathbf{ch}}(N_1,\dots N_{i-1},M_1,\dots,M_n,N_{i+1},\dots,N_m;K\otimes\mathcal{D}_X\otimes\mathcal{D}_X)\\
\cong\ &i_{2*}\mathcal{P}_{\mathbf{ch}}(N_1,\dots N_{i-1},M_1,\dots,M_n,N_{i+1},\dots,N_m,K\otimes\mathcal{D}_X).
\end{aligned}$$

By substituting M instead of all N_i, M_j, K, we get the desired insertion map

$$\circ_i : U_M(n)\boxtimes U_M(m)\to i_{2*}U_M(n+m-1).$$

2.5.7.

Similarly, we have insertion maps

$$\circ_i : U_M(n)\otimes\mathcal{P}_{\mathbf{ch}_M}(m)\to U_M(n+m-1),$$

and

$$\circ_i : P_{\mathbf{ch}_M}(n)\otimes U_M(m)\to U_M(n+m-1).$$

2.5.8.

Let $\mathcal{O}$ be a differential graded operad. Set

$$\mathfrak{g}_{\mathcal{O},n}:=\mathcal{O}(n)^{S_n},$$

and $\mathfrak{g}_{\mathcal{O}}=\oplus_n\mathfrak{g}_{\mathcal{O},n}[1-n]$.

Let $p_n:\mathcal{O}(n)\to\mathfrak{g}_{\mathcal{O},n}$ be the natural projection, which is the symmetrization map. Define the brace $(x,y)\mapsto x\{y\}$, $\mathfrak{g}_{\mathcal{O},n}\otimes\mathfrak{g}_{\mathcal{O},m}\to\mathfrak{g}_{\mathcal{O},n+m-1}$ as follows.

$$x\{y\}=np_n(\circ_1(x,y)) \tag{2}$$

and the bracket

$$[x,y]=x\{y\}-(-1)^{|x||y|}y\{x\}. \tag{3}$$

We see that $[,]$ is a Lie bracket. Thus, $\mathfrak{g}_{\mathcal{O}}$ is a differential graded Lie algebra. For an operad $\mathcal{O}$ denote by $\mathcal{O}'$ the shifted operad such that the structure of an $\mathcal{O}'$-algebra on a complex V is equivalent to the structure of an $\mathcal{O}$-algebra on a complex $V[1]$. Thus, $\mathcal{O}'(n)=\mathcal{O}(n)\otimes\epsilon_n[1-n]$, where ϵ_n is the sign representation of S_n.

Let $\mathcal{O}$ be an operad of vector spaces. The set of Maurer-Cartan elements of $\mathfrak{g}_{\mathcal{O}'}$ is in 1-1 correspondence with maps of operads $\mathbf{lie}\to\mathcal{O}$.

Assume that $O(1)$ is a nilpotent algebra ($x^n=0$ for any $x\in\mathcal{O}(1)$). Let A be $O(1)$ with adjoined unit and let $A^\times$ be the group of invertible elements. $A^\times$ acts on $\mathcal{O}$ by automorphisms. Therefore, $A^\times$ acts on the set of maps $\mathbf{lie}\to\mathcal{O}$. The groupoid of this action is isomorphic to the Deligne groupoid of $\mathfrak{g}_{\mathcal{O}'}$.

2.5.9.

Similarly, let $\mathcal{A}$ be a *-operad. Then formula 3 defines a Lie-* algebra $\mathfrak{g}_{\mathcal{A}}$. We have natural action of a usual pro-Lie algebra $\mathfrak{g}_{\mathcal{P}_{\mathbf{ch}}(M)}$ on a pro *-Lie algebra $\mathfrak{g}_{U_M}$ by derivations. The chiral bracket $b \in \mathfrak{g}^1_{\mathcal{P}_{\mathbf{ch}}(M)}$ satisfies $[b,b]=0$. Therefore, the bracket with b defines a differential on $\mathfrak{g}_{U_M}$. Denote this differential graded *-Lie algebra by $\mathfrak{d}_M$.

2.5.10.

To avoid using derived functors, we will slightly modify $\mathfrak{d}_M$. Recall that $M = \omega_X \oplus N$, where N is free. Let

$$\mathcal{P}^{\mathbf{red}}_{\mathbf{ch}}(M,\dots,M;M\otimes\mathcal{D}_X) \subset \mathcal{P}_{\mathbf{ch}}(M,\dots,M;M\otimes\mathcal{D}_X)$$

be the subset of all operations vanishing under all restrictions

$$\mathcal{P}_{\mathbf{ch}}(M,\dots,M;M\otimes\mathcal{D}_X) \to \mathcal{P}_{\mathbf{ch}}(M,\dots,M,\omega_X,M,\dots,M;M\otimes\mathcal{D}_X).$$

Let $\mathbf{def}_M \subset \mathfrak{d}_M$ be the submodule such that

$$\mathbf{def}_M = \oplus_n(\mathcal{P}^{\mathbf{red}}_{\mathbf{ch}}(M,\dots,M;M\otimes\mathcal{D}_X)\otimes\epsilon_n)^{S_n}[1-n].$$

We see that $\mathbf{def}_M$ is a *-Lie differential subalgebra of $\mathfrak{d}_M$.

2.5.11.

Proposition 2.4 *The functors G_M and $F_{\mathbf{def}_M}$ are canonically isomorphic.*

2.6. Example

Let K be a free left $\mathcal{D}_X$-module. Let $T^iK = K^{\otimes_{\mathcal{O}_X} i}$. The symmetric group S_i acts on the $\mathcal{D}_X$-module T^iK; let $S^iK = (T^iK)^{S_i}$ be the submodule of invariants and $SK = \oplus_{i=0}^{\infty} S^iK$. SK is naturally a free commutative $\mathcal{D}_X$-algebra and, hence, $SK^r := SK\otimes\omega_X$ is a chiral algebra. We will compute the cohomology of the $\mathcal{D}_X$-module $\mathbf{def}_{SK^r}$. Let $S_0K = \oplus_{n=1}^{\infty} S^iK$. We have:

$$\mathbf{def}_{SK^r} = \oplus_n(P_{\mathbf{ch}}(S_0K^r[1],\dots,S_0K^r[1];SK^r\otimes\mathcal{D}_X)[1])^{S_n}.$$

On the other hand, denote by $\Omega := SK\otimes K$. Consider Ω as an SK-$\mathcal{D}_X$-module of differentials of SK. We have the de Rham differential $D : S_0K \to \Omega$. We have a through map

$$\begin{aligned} c_n : P_{\mathbf{ch}}(K[1]^r,\dots,K[1]^r,SK[1]^r)^{S_n} &\cong P^{SK}_{\mathbf{ch}}(\Omega[1]^r,\dots,\Omega[1]^r,SK[1]^r)^{S_n} \\ &\xrightarrow{D} P_{\mathbf{ch}}(S_0K[1]^r,\dots,S_0K[1]^r,SK[1]^r)^{S_n}, \end{aligned}$$

where $P^{SK}_{\mathbf{ch}}$ stands for SK-linear chiral operations. Denote by the same letter the induced map

$$c_n : P_{\mathbf{ch}}(K[1]^r,\dots,K[1]^r,SK[1]^r)^{S_n} \to \mathbf{def}_{SK}.$$

Proposition 2.5 *(1)* $dc_n = 0$*;*
(2) c_n *induces an isomorphism*

$$P_{\mathbf{ch}}(K[1]^r, \dots, K[1]^r, SK[1]^r)^{S_n} \to H^{n-1}(\mathbf{def}_{SK})[1-n].$$

2.6.1.

For a chiral algebra M denote by H_M the graded Lie algebra of cohomology of $\mathbf{def}_M$.

2.6.2.

Assume that K is finitely generated. Let

$$K^\vee = \hom(K, \mathcal{D}_X) \otimes (\omega_X)^{-1}$$

be the dual module. Then

$$H_{SK^r} \cong \oplus_n (P_*(K^r, \dots, K^r; SK^r \otimes \mathcal{D}_X)\varepsilon_n)^{S_n}[1-n] = \oplus_n (\wedge^n K^\vee \otimes_{\mathcal{O}_X} SK)^r[1-n].$$

2.6.3.

We will postpone the calculation of the *-Lie bracket on H_{SK^r} until we show in the next section that H_M has in fact a richer structure.

3. Algebraic structure on the cohomology of the deformation pro-*-Lie algebra

We will keep the agreements and the notations from 2.2.1..

3.1. Cup product

We will define a chiral operation $\cup \in P_{\mathbf{ch}\ \mathbf{def}_{M[-1]}}(2)$ and then we will study the induced map on cohomology.

3.1.1.

Recall that

$$\mathbf{def}_M[-1] \cong \oplus_n (a_n)^{S_n},$$

where

$$a_n = P_{\mathbf{ch}}(N[1], \dots, N[1]; M \otimes \mathcal{D}_X).$$

Let $i_n : X \to X^n$ be the diagonal embedding and let $p^i : X^n \to X$ $U_n \subset X^n$ be the complement to the union of all pairwise diagonals $p^i x = p^j x$ and $j_n : U_n \to X_n$ be the open embedding. Let $U_{n,m} \subset X^{n+m}$ be the complement to the diagonals $p^i x = p^j x$, where $1 \le i \le n$, $n+1 \le j \le n+m$ and $j_{nm} : U_{nm} \to X^{n+m}$ be the embedding

Compute

$$\begin{aligned} j_{2*}j_2^*(a_n \boxtimes a_m) &\cong \hom(j_{n*}j_n^*(N[1]^{\boxtimes n}) \boxtimes j_{m*}j_m^*(N[1]^{\boxtimes m}), \\ &\quad (i_n \times i_m)_*(j_{2*}j_2^*(M \otimes \mathcal{D}_X \boxtimes M \otimes \mathcal{D}_X))) \\ &\cong \hom(j_{n*}j_n^*(N[1]^{\boxtimes n}) \boxtimes j_{m*}j_m^*(N[1]^{\boxtimes m}) \otimes j_*\mathcal{O}(U_{n,m}), \\ &\quad (i_n \times i_m)_*(j_{2*}j_2^*(M \otimes \mathcal{D}_X \boxtimes M \otimes \mathcal{D}_X) \otimes j_*\mathcal{O}(U_{n,m}))) \\ &\cong \hom(j_{n+m*}j_{n+m}^*(N[1])^{\boxtimes n+m}, (i_n \times i_m)_*(j_{2*}j_2^*(M \boxtimes M)) \otimes \mathcal{D}_{X\times X})). \end{aligned}$$

Taking the composition with the chiral operation on M, we obtain a chiral operation

$$j_{2*}j_2^*(a_n \boxtimes a_m) \to \hom(j_{n+m*}j_{n+m}^*(N[1])^{\boxtimes n+m}, i_{n+m*}(M \otimes \mathcal{D}_X \otimes \mathcal{D}_X)) \cong i_{2*}a_{n+m},$$

which induces a chiral operation from $P_{\mathbf{ch}}(a_n^{S_n}, a_m^{S_m}; a_{n+m}^{S_{n+m}})$ and, hence, an operation $\cup \in (P_{\mathbf{ch}}(\mathbf{def}_M[-1], \mathbf{def}_M[-1]; \mathbf{def}_M[-1]))^{S_2}$.

3.1.2.

To investigate the properties of this operation, consider the brace *-operation $\cdot\{\} \in P_*(\mathbf{def}_M, \mathbf{def}_M; \mathbf{def}_M)$ defined by formula (2). Let

$$r : P_{\mathbf{ch}}(A_1, A_2; A_3) \to P_*(A_1, A_2; A_3)$$

be the natural map

Proposition 3.1 $d(\cdot\{\}) = r(\cup)$.

Let $\cup_h$ be the induced operation on $H_M[-1]$. The above proposition implies that $r(\cup_h) = 0$. In virtue of exact sequence

$$0 \to \hom((A_1)^l \otimes (A_2)^l, (A_3)^l) \to P_{\mathbf{ch}}(A_1, A_2; A_3) \xrightarrow{r} P_*(A_1, A_2; A_3),$$

$\cup_h$ defines a $\mathcal{D}_X$-commutative product $H_M[-1] \otimes H_M[-1] \to H_M[-1]$, denoted by the same letter.

3.1.3.

Proposition 3.2 $\cup_h$ *is associative.*

3.1.4. Leibnitz rule

We are going to establish a relation between $\cup$ and $\cdot\{\}$. This relation is similar to the one of coisson algebras. Our exposition will mimic the definition of coisson algebras from [1].

3.1.5.

Let A_i be right $\mathcal{D}_X$-modules. Write $A_1 \overset{!}{\otimes} A_2 := (A_1^l \otimes A_2^l)^r$; $P_!(A_1, A_2; A_3) := \hom(A_1 \overset{!}{\otimes} A_2, A_3)$. We have $(A \overset{!}{\otimes} B) = i_2^*(B \boxtimes C)$;

$$i_{2*}(A \overset{!}{\otimes} B) \to i_{2*}A \otimes p_2^*(B^l).$$

We have a map

$$c : P_*(A_1, A_2; B) \otimes P_!(B, C; D) \to P_*(A_1, A_2 \overset{!}{\otimes} C; D)$$

defined as follows. Let $u : A_1 \boxtimes A_2 \to i_{2*}B$ and $m : B \overset{!}{\otimes} C \to D$. Put

$$c(u, v) : A_1 \boxtimes (A_2 \overset{!}{\otimes} C) \cong (A_1 \boxtimes A_2) \otimes p_2^*(C^l) \to i_{2*}B \otimes p_2^*(C^l) \cong i_{2*}(B \overset{!}{\otimes} C) \to i_{2*}D.$$

3.1.6.

Denote

$$e = c([], \cup_h) \in P_*(H_M, H_M \overset{!}{\otimes} H_M; H_M).$$

Let $T : H_M \overset{!}{\otimes} H_M \to H_M \overset{!}{\otimes} H_M$ be the action of symmetric group and let e^T be the composition with T. Let $f \in P_*(H_M, H_M \overset{!}{\otimes} H_M; H_M)$ be defined by:

$$H_M \boxtimes (H_M \overset{!}{\otimes} H_M) \overset{\cup}{\to} H_M \boxtimes H_M \overset{[]}{\to} i_* H_M.$$

Proposition 3.3 *We have* $f = e + e^T$.

In other words, the cup product and the bracket satisfy the Leibnitz identity.

3.1.7.

We see that H_M has a pro-*-Lie bracket, $(H_M)^l[1]$ has a commutative $\mathcal{D}_X$-algebra structure, and these structures satisfy the Leibnitz identity. Call this structure *a c-Gerstenhaber algebra* structure. Thus, our findings can be summarized as follows.

Theorem 3.4 *The cohomology of the deformation pro-*-Lie algebra of a chiral algebra is naturally a pro-c-Gerstenhaber algebra.*

3.2. Example $M = (SK)^r$

We come back to our example 2.6.. For simplicity assume K is finitely generated free $\mathcal{D}_X$-module. We have seen in this case that

$$(H_M)^l[-1] \cong \oplus_i \wedge^i K^\vee \otimes S^K[-i] \cong S(K^\vee[-1] \oplus K).$$

Proposition 3.5 *The cup product on H_M coincides with the natural one on the symmetric power algebra.*

3.2.1.

To describe the bracket it suffices to define it on the submodule of generators $G = (K^\vee[-1] \oplus K)^r$. Define $[] \in P_*(G, G; H_M)$ to be zero when restricted onto $K^r \boxtimes K^r$ and $K^{\vee r}[-1] \boxtimes K^{\vee r}[-1]$. Restriction onto $K \boxtimes K^\vee[-1]$ takes values in $\omega_X \subset H_M$ and is given by the canonic $*$-pairing from [1]

$$(K^\vee \boxtimes K)^r \to i_{2*}\omega_X.$$

Recall the definition. We have $K^{\vee r} = \hom(K^r, \mathcal{D}_X \otimes \omega_X)$. For open $U, V \subset X$ we have the composition map

$$K(U) \otimes K^\vee(V) \to \mathcal{D}_X \otimes \omega_X(U \cap V) \cong i_{2*}\omega_X(U \times V)$$

which defines the pairing. This uniquely defines the *-Lie bracket.

4. Formality Conjecture

Following the logic of Kontsevich's formality theorem, one can formulate a formality conjecture in this situation.

4.1. Quasi-isomorphisms

A map $f : \mathfrak{g} \to \mathfrak{h}$ of differential graded pro-*-Lie algebras is called quasi-isomorphism if it induces an isomorphism on cohomology. Call a pro-*-Lie algebra *perfect* if such is its underlying complex of pro-vector spaces. The morphism f is called *perfect quasi-isomorphism* if it is a quasi-isomorphism and both $\mathfrak{g}$ and $\mathfrak{h}$ are perfect.

Two perfect pro-*-lie algebras are called perfectly quasi-isomorphic if there exists a chain of perfect quasi-isomorphisms connecting $\mathfrak{g}$ and $\mathfrak{h}$.

Conjecture 4.1 $\mathbf{def}_{SK}$ *and* H_{SK} *are perfectly quasi-isomorphic.*

The importance of this conjecture can be seen from the following theorem:

Theorem 4.2 *Any chain of perfect quasi-isomorphisms between* $\mathbf{def}_{SK^r}$ *and* H_{SK^r} *establishes a bijection between the set of isomorphism classes of* A*-linear coisson brackets on* $SK^r \otimes A$ *which vanish modulo the maximal ideal* $\mathcal{M}_A$ *and the set of isomorphism classes of all deformations of the chiral algebra* SK^r *over* A.

References

[1] A. Beilinson, V. Drinfeld, Chiral Algebras.
[2] W. Goldman, J. Millson, Deformations of Flat Bundles over Kähler Manifolds, Geometry and Topology, 129–145, Lect. Notes in Pure and Applied Math. **105** Dekker NY (1987).
[3] M. Kontsevich, Quantization of Poisson Manifolds, preprint.

Section 3. Number Theory

J. W. Cogdell, I. I. Piatetski-Shapiro: *Converse Theorems, Functoriality, and Applications to Number Theory* 119

H. Cohen: *Constructing and Counting Number Fields* 129

Jean-Marc Fontaine: *Analyse p-adique et Représentations Galoisiennes* 139

A. Huber, G. Kings: *Equivariant Bloch-Kato Conjecture and Non-abelian Iwasawa Main Conjecture* 149

Kazuya Kato: *Tamagawa Number Conjecture for zeta Values* 163

Stephen S. Kudla: *Derivatives of Eisenstein Series and Arithmetic Geometry* 173

Barry Mazur, Karl Rubin: *Elliptic Curves and Class Field Theory* 185

Emmanuel Ullmo: *Théorie Ergodique et Géométrie Arithmétique* 197

Trevor D. Wooley: *Diophantine Methods for Exponential Sums, and Exponential Sums for Diophantine Problems* 207

ICM 2002 · Vol. II · 119–128

Converse Theorems, Functoriality, and Applications to Number Theory

J. W. Cogdell* **I. I. Piatetski-Shapiro†**

Abstract

There has been a recent coming together of the Converse Theorem for GL_n and the Langlands-Shahidi method of controlling the analytic properties of automorphic L-functions which has allowed us to establish a number of new cases of functoriality, or the lifting of automorphic forms. In this article we would like to present the current state of the Converse Theorem and outline the method one uses to apply the Converse Theorem to obtain liftings. We will then turn to an exposition of the new liftings and some of their applications.

2000 Mathematics Subject Classification: 11F70, 22E55.
Keywords and Phrases: Automorphic forms, L-functions, Converse theorems, Functoriality.

1. Introduction

Converse Theorems traditionally have provided a way to characterize Dirichlet series associated to modular forms in terms of their analytic properties. Most familiar are the Converse Theorems of Hecke and Weil. Hecke first proved that L-functions associated to modular forms enjoyed "nice" analytic properties and then proved "Conversely" that these analytic properties in fact characterized modular L-functions. Weil extended this Converse Theorem to L-functions of modular forms with level.

In their modern formulation, Converse Theorems are stated in terms of automorphic representations of $\mathrm{GL}_n(\mathbb{A})$ instead of modular forms. Jacquet, Piatetski-Shapiro, and Shalika have proved that the L-functions associated to automorphic representations of $\mathrm{GL}_n(\mathbb{A})$ have nice analytic properties via integral representations similar to those of Hecke. The relevant "nice" properties are: analytic continuation, boundedness in vertical strips, and functional equation. Converse Theorems in this context invert these integral representations. They give a criterion for an irreducible

*Department of Mathematics, Oklahoma State University, Stillwater, OK 74078, USA. E-mail: cogdell@math.okstate.edu

†Department of Mathematics, Yale University, New Haven, CT 06520, USA, and School of Mathematics, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: ilya@math.yale.edu

admissible representation Π of $\mathrm{GL}_n(\mathbb{A})$ to be automorphic and cuspidal in terms of the analytic properties of Rankin-Selberg convolution L-functions $L(s, \Pi \times \pi')$ of Π twisted by cuspidal representations π' of $\mathrm{GL}_m(\mathbb{A})$ of smaller rank groups.

To use Converse Theorems for applications, proving that certain objects are automorphic, one must be able to show that certain L-functions are "nice". However, essentially the only way to show that an L-function is nice is to have it associated to an automorphic form. Hence the most natural applications of Converse Theorems are to functoriality, or the lifting of automorphic forms, to GL_n. More explicit number theoretic applications then come as consequences of these liftings.

Recently there have been several applications of Converse Theorems to establishing functorialities. These have been possible thanks to the recent advances in the Langlands-Shahidi method of analysing the analytic properties of general automorphic L-functions, due to Shahidi and his collaborators [21]. By combining our Converse Theorems with their control of the analytic properties of L-functions many new examples of functorial liftings to GL_n have been established. These are described in Section 4 below. As one number theoretic consequence of these liftings Kim and Shahidi have been able to establish the best general estimates over a number field towards the Ramamujan-Selberg conjectures for GL_2, which in turn have already had other applications.

2. Converse Theorems for GL_n

Let k be a global field, $\mathbb{A}$ its adele ring, and ψ a fixed non-trivial (continuous) additive character of $\mathbb{A}$ which is trivial on k. We will take $n \geq 3$ to be an integer.

To state these Converse Theorems, we begin with an irreducible admissible representation Π of $\mathrm{GL}_n(\mathbb{A})$. It has a decomposition $\Pi = \otimes' \Pi_v$, where Π_v is an irreducible admissible representation of $\mathrm{GL}_n(k_v)$. By the local theory of Jacquet, Piatetski-Shapiro, and Shalika [9, 11] to each Π_v is associated a local L-function $L(s, \Pi_v)$ and a local ε-factor $\varepsilon(s, \Pi_v, \psi_v)$. Hence formally we can form

$$L(s, \Pi) = \prod L(s, \Pi_v) \qquad \text{and} \qquad \varepsilon(s, \Pi, \psi) = \prod \varepsilon(s, \Pi_v, \psi_v).$$

We will always assume the following two things about Π:

(1) $L(s, \Pi)$ converges in some half plane $Re(s) >> 0$,
(2) the central character ω_Π of Π is automorphic, that is, invariant under $k^\times$.

Under these assumptions, $\varepsilon(s, \Pi, \psi) = \varepsilon(s, \Pi)$ is independent of our choice of ψ [4].

As in Weil's case, our Converse Theorems will involve twists but now by cuspidal automorphic representations of $\mathrm{GL}_m(\mathbb{A})$ for certain m. For convenience, let us set $\mathcal{A}(m)$ to be the set of automorphic representations of $\mathrm{GL}_m(\mathbb{A})$, $\mathcal{A}_0(m)$ the set of (irreducible) cuspidal automorphic representations of $\mathrm{GL}_m(\mathbb{A})$, and $\mathcal{T}(m) = \bigcup_{d=1}^{m} \mathcal{A}_0(d)$. If S is a finite set of places, we will let $\mathcal{T}^S(m)$ denote the subset of representations $\pi \in \mathcal{T}$ with local components π_v unramified at all places $v \in S$ and let $\mathcal{T}_S(m)$ denote those π which are unramified for all $v \notin S$.

Let $\pi' = \otimes' \pi'_v$ be a cuspidal representation of $\mathrm{GL}_m(\mathbb{A})$ with $m < n$. Then again we can formally define

$$L(s, \Pi \times \pi') = \prod L(s, \Pi_v \times \pi'_v) \qquad \text{and} \qquad \varepsilon(s, \Pi \times \pi') = \prod \varepsilon(s, \Pi_v \times \pi'_v, \psi_v)$$

since the local factors make sense whether Π is automorphic or not. A consequence of (1) and (2) above and the cuspidality of π' is that both $L(s, \Pi \times \pi')$ and $L(s, \widetilde{\Pi} \times \widetilde{\pi'})$ converge absolutely for $Re(s) >> 0$, where $\widetilde{\Pi}$ and $\widetilde{\pi'}$ are the contragredient representations, and that $\varepsilon(s, \Pi \times \pi')$ is independent of the choice of ψ.

We say that $L(s, \Pi \times \pi')$ is *nice* if it satisfies the same analytic properties it would if Π were cuspidal, i.e.,

1. $L(s, \Pi \times \pi')$ and $L(s, \widetilde{\Pi} \times \widetilde{\pi'})$ have continuations to *entire* functions of s,
2. these entire continuations are *bounded in vertical strips* of finite width,
3. they satisfy the standard *functional equation*

$$L(s, \Pi \times \pi') = \varepsilon(s, \Pi \times \pi') L(1 - s, \widetilde{\Pi} \times \widetilde{\pi'}).$$

The basic converse theorem for GL_n is the following.

Theorem 1. [6] *Let Π be an irreducible admissible representation of $\mathrm{GL}_n(\mathbb{A})$ as above. Let S be a finite set of finite places. Suppose that $L(s, \Pi \times \pi')$ is nice for all $\pi' \in \mathcal{T}^S(n-2)$. Then Π is quasi-automorphic in the sense that there is an automorphic representation Π' such that $\Pi_v \simeq \Pi'_v$ for all $v \notin S$. If S is empty, then in fact Π is a cuspidal automorphic representation of $\mathrm{GL}_n(\mathbb{A})$.*

It is this version of the Converse Theorem that has been used in conjunction with the Langlands-Shahidi method of controlling analytic properties of L-functions in the new examples of functoriality explained below.

Theorem 2. [4] *Let Π be an irreducible admissible representation of $\mathrm{GL}_n(\mathbb{A})$ as above. Let S be a non-empty finite set of places, containing S_∞, such that the class number of the ring $\mathfrak{o}_S$ of S-integers is one. Suppose that $L(s, \Pi \times \pi')$ is nice for all $\pi' \in \mathcal{T}_S(n-1)$. Then Π is quasi-automorphic in the sense that there is an automorphic representation Π' such that $\Pi_v \simeq \Pi'_v$ for all $v \in S$ and all $v \notin S$ such that both Π_v and Π'_v are unramified.*

This version of the Converse Theorem was specifically designed to investigate functoriality in the cases where one controls the L-functions by means of integral representations where it is expected to be more difficult to control twists.

The proof of Theorem 1 with S empty and $n-2$ replaced by $n-1$ essentially follows the lead of Hecke, Weil, and Jacquet-Langlands. It is based on the integral representations of L-functions, Fourier expansions, Mellin inversion, and finally a use of the weak form of Langlands spectral theory. For Theorems 1 and 2 where we have restricted our twists either by ramification or rank we must impose certain local conditions to compensate for our limited twists. For Theorem 1 are a finite number of local conditions and for Theorem 2 an infinite number of local conditions. We must then work around these by using results on generation of congruence subgroups and either weak approximation (Theorem 1) or strong approximation (Theorem 2).

As for our expectations of what form the Converse Theorem may take in the future, we refer the reader to the last section of [6].

3. Functoriality via the Converse Theorem

In order to apply these theorems, one must be able to control the analytic properties of the L-function. However the only way we have of controlling global L-functions is to associate them to automorphic forms or representations. A minute's thought will then convince one that the primary application of these results will be to the lifting of automorphic representations from some group H to GL_n.

Suppose that H is a reductive group over k. For simplicity of exposition we will assume throughout that H is split and deal only with the connected component of its L-group, which we will (by abuse of notation) denote by ${}^L\mathrm{H}$ [1]. Let $\pi = \otimes'\pi_v$ be a cuspidal automorphic representation of H and ρ a complex representation of ${}^L\mathrm{H}$. To this situation Langlands has associated an L-function $L(s,\pi,\rho)$ [1]. Let us assume that ρ maps ${}^L\mathrm{H}$ to $\mathrm{GL}_n(\mathbb{C})$. Then by Langlands' general Principle of Functoriality to π should be associated an automorphic representation Π of $\mathrm{GL}_n(\mathbb{A})$ satisfying $L(s,\Pi) = L(s,\pi,\rho)$, $\varepsilon(s,\Pi) = \varepsilon(s,\pi,\rho)$, with similar equalities locally and for the twisted versions [1]. Using the Converse Theorem to establish such liftings involves three steps: construction of a candidate lift, verification that the twisted L-functions are "nice", and application of the appropriate Converse Theorem.

1. *Construction of a candidate lift*: We construct a candidate lift $\Pi = \otimes'\Pi_v$ on $\mathrm{GL}_n(\mathbb{A})$ place by place. We can see what Π_v should be at almost all places. Since we have the arithmetic Langlands (or Hecke-Frobenius) parameterization of representations π_v of $\mathrm{H}(k_v)$ for all archimedean places and those non-archimedean places where the representations are unramified [1], we can use these to associate to π_v and the map $\rho_v : {}^L\mathrm{H}_v \to {}^L\mathrm{H} \to \mathrm{GL}_n(\mathbb{C})$ a representation Π_v of $\mathrm{GL}_n(k_v)$. This correspondence preserves local L- and ε-factors

$$L(s,\Pi_v) = L(s,\pi_v,\rho_v) \qquad \text{and} \qquad \varepsilon(s,\Pi_v,\psi_v) = \varepsilon(s,\pi_v,\rho_v,\psi_v)$$

along with the twisted versions. If H happens to be GL_m or a related group then we in principle know how to associate the representation Π_v at all places now that the local Langlands conjecture has been solved for GL_m. For other situations, we may not know what Π_v should be at the ramified places. We will return to this difficulty momentarily and show how one can work around this with the use of a highly ramified twist. But for now, let us assume we can finesse this local problem and arrive at a global representation $\Pi = \otimes'\Pi_v$ such that

$$L(s,\Pi) = \prod L(s,\Pi_v) = \prod L(s,\pi_v,\rho_v) = L(s,\pi,\rho)$$

and similarly $\varepsilon(s,\Pi) = \varepsilon(s,\pi,\rho)$ with similar equalities for the twisted versions. Π should then be the Langlands lifting of π to $\mathrm{GL}_n(\mathbb{A})$ associated to ρ.

2. *Analytic properties of global L-functions*: For simplicity of exposition, let us now assume that ρ is simply a standard embedding of ${}^L\mathrm{H}$ into $\mathrm{GL}_n(\mathbb{C})$, such as will be the case if we consider H to be a split classical group, so that $L(s,\pi,\rho) = L(s,\pi)$ is the standard L-function of π. We have our candidate Π for the lift of π to GL_n from above. To be able to assert that the Π which we constructed place by place is automorphic, we will apply a Converse Theorem. To do so we must control the twisted L-functions $L(s,\Pi\times\pi') = L(s,\pi\times\pi')$ for $\pi' \in \mathcal{T}$ with an appropriate

twisting set $\mathcal{T}$ from Theorem 1 or 2. In the examples presented below, we have used Theorem 1 above and the analytic control of $L(s, \pi \times \pi')$ achieved by the so-called Langlands-Shahidi method of analyzing the L-functions through the Fourier coefficients of Eisenstein series [21]. Currently this requires us to take k to be a number field. The *functional equation* $L(s, \pi \times \pi') = \varepsilon(s, \pi \times \pi')L(1-s, \tilde{\pi} \times \tilde{\pi}')$ has been proved in wide generality by Shahidi [18]. The *boundedness in vertical strips* has been proved in close to the same generality by Gelbart and Shahidi [7]. As for the entire continuation of $L(s, \pi \times \pi')$, a moments thought will tell you that one should not always expect a cuspidal representation of $\mathrm{H}(\mathbb{A})$ to necessarily lift to a cuspidal representation of $\mathrm{GL}_n(\mathbb{A})$. Hence it is unreasonable to expect all $L(s, \pi \times \pi')$ to be entire. We had previously understood how to work around this difficulty from the point of view of integral representations by again using a highly ramified twist. Kim realized that one could also control the entirety of these twisted L-functions in the context of the Langlands-Shahidi method by using a highly ramified twist. We will return to this below. Thus in a fairly general context one has that $L(s, \pi \times \pi')$ is *entire* for π' in a suitably modified twisting set $\mathcal{T}'$.

3. *Application of the Converse Theorem*: Once we have that $L(s, \pi \times \pi')$ is nice for a suitable twisting set $\mathcal{T}'$ then from the equalities

$$L(s, \Pi \times \pi') = L(s, \pi \times \pi') \qquad \text{and} \qquad \varepsilon(s, \Pi \times \pi') = \varepsilon(s, \pi \times \pi')$$

we see that the $L(s, \Pi \times \pi')$ are nice and then we can apply our Converse Theorems to conclude that Π is either cuspidal automorphic or at least that there is an automorphic Π' such that $\Pi_v = \Pi'_v$ at almost all places. This then effects the (possibly weak) automorphic lift of π to Π or Π'.

4. *Highly ramified twists*: As we have indicated above, there are both local and global problems that can be finessed by an appropriate use of a highly ramified twist. This is based on the following simple observation.

Observation. *Let Π be as in Theorem 1 or 2. Suppose that η is a fixed character of $k^\times \backslash \mathbb{A}^\times$. Suppose that $L(s, \Pi \times \pi')$ is nice for all $\pi' \in \mathcal{T}' = \mathcal{T} \otimes \eta$, where $\mathcal{T}$ is either of the twisting sets of Theorem 1 or 2. Then Π is quasi-automorphic as in those theorems.*

The only thing to observe is that if $\pi' \in \mathcal{T}$ then $L(s, \Pi \times (\pi' \otimes \eta)) = L(s, (\Pi \otimes \eta) \times \pi')$ so that applying the Converse Theorem for Π with twisting set $\mathcal{T} \otimes \eta$ is equivalent to applying the Converse Theorem for $\Pi \otimes \eta$ with the twisting set $\mathcal{T}$. So, by either Theorem 1 or 2, whichever is appropriate, $\Pi \otimes \eta$ is quasi-automorphic and hence Π is as well.

If we now begin with π automorphic on $\mathrm{H}(\mathbb{A})$, we will take T to be the set of finite places where π_v is ramified. For applying Theorem 1 we want $S = T$ and for Theorem 2 we would want $S \cap T = \emptyset$. We will now take η to be highly ramified at all places $v \in T$, so that at $v \in T$ our twisting representations are all locally of the form (unramified principal series)$\otimes$(highly ramified character).

In order to finesse the lack of knowledge of an appropriate local lift, we need to know the following two local facts about the local theory of L-functions for H.

Multiplicativity of γ-factors. *If $\pi'_v = Ind(\pi'_{1,v} \otimes \pi'_{2,v})$, with $\pi'_{i,v}$ and irreducible admissible representation of $\mathrm{GL}_{r_i}(k_v)$, then we have $\gamma(s, \pi_v \times \pi'_v, \psi_v) = \gamma(s, \pi_v \times \pi'_{1,v}, \psi_v)\gamma(s, \pi_v \times \pi'_{2,v}, \psi_v)$.*

Stability of γ-factors. *If $\pi_{1,v}$ and $\pi_{2,v}$ are two irreducible admissible representations of $\mathrm{H}(k_v)$ with the same central character, then for every sufficiently highly ramified character η_v of $GL_1(k_v)$ we have $\gamma(s, \pi_{1,v} \times \eta_v, \psi_v) = \gamma(s, \pi_{2,v} \times \eta_v, \psi_v)$.*

Both of these facts are known for GL_n, the multiplicativity being found in [9] and the stability in [10]. Multiplicativity in a fairly wide generality useful for applications has been established by Shahidi [19]. Stability is in a more primitive state at the moment, but Shahidi has begun to establish the necessary results in a general context in [20].

To utilize these local results, what one now does is the following. At the places where π_v is ramified, choose Π_v to be arbitrary, except that it should have the same central character as π_v. This is both to guarantee that the central character of Π is the same as that of π and hence automorphic and to guarantee that the stable forms of the γ-factors for π_v and Π_v agree. Now form $\Pi = \otimes' \Pi_v$. Choose our character η so that at the places $v \in T$ we have that the L- and γ-factors for both $\pi_v \otimes \eta_v$ and $\Pi_v \otimes \eta_v$ are in their stable form and agree. We then twist by $\mathcal{T}' = \mathcal{T} \otimes \eta$ for this *fixed* character η. If $\pi' \in \mathcal{T}'$, then for $v \in T$, π'_v is of the form $\pi'_v = Ind(|\ |^{s_1} \otimes \cdots \otimes |\ |^{s_m}) \otimes \eta_v$. So at the places $v \in T$, applying both multilplcativity and stability, we have

$$\begin{aligned}\gamma(s, \pi_v \times \pi'_v, \psi_v) &= \prod \gamma(s + s_i, \pi_v \otimes \eta_v, \psi_v) \\ &= \prod \gamma(s + s_i, \Pi_v \otimes \eta_v, \psi_v) = \gamma(s, \Pi_v \times \pi'_v, \psi_v)\end{aligned}$$

from which one deduces a similar equality for the L- and ε-factors. From this it will then follow that globally we will have $L(s, \pi \times \pi') = L(s, \Pi \times \pi')$ for all $\pi' \in \mathcal{T}'$ with similar equalities for the ε-factors. This then completes Step 1.

To complete our use of the highly ramified twist, we must return to the question of whether $L(s, \pi \times \pi')$ can be made entire. In analysing L-functions via the Langlands-Shahidi method, the poles of the L-function are controlled by those of an Eisenstein series. In general, the inducing data for the Eisenstein series must satisfy a type of self-contragredience for there to be poles. The important observation of Kim is that one can use a highly ramified twist to destroy this self-contragredience at one place, which suffices, and hence eliminate poles. The precise condition will depend on the individual construction. A more detailed explanation of this can be found in Shahidi's article [21]. This completes Step 2 above.

4. New examples of functoriality

Now take k to be a number field. There has been much progress recently in utilizing the method described above to establish global liftings from split groups H over k to an appropriate GL_n. Among them are the following.

1. *Classical groups.* Take H to be a split classical group over k, more specifically, the split form of either SO_{2n+1}, Sp_{2n}, or SO_{2n}. The the L-groups ${}^L\mathrm{H}$ are then $\mathrm{Sp}_{2n}(\mathbb{C})$, $\mathrm{SO}_{2n+1}(\mathbb{C})$, or $\mathrm{SO}_{2n}(\mathbb{C})$ and there are natural embeddings into the general linear group $\mathrm{GL}_{2n}(\mathbb{C})$, $\mathrm{GL}_{2n+1}(\mathbb{C})$, or $\mathrm{GL}_{2n}(\mathbb{C})$ respectively. Associated to each there should be a lifting of admissible or automorphic representations from

$\mathrm{H}(\mathbb{A})$ to the appropriate $\mathrm{GL}_N(\mathbb{A})$. The first lifting that resulted from the combination of the Converse Theorem and the Langlands-Shahidi method of controlling automorphic L-functions was the weak lift for generic cuspidal representations from SO_{2n+1} to GL_{2n} over a number field k obtained with Kim and Shahidi [2]. We can now extend this to the following result.

Theorem. [2, 3] *Let* H *be a split classical group over* k *as above and* π *a globally generic cuspidal representation of* $\mathrm{H}(\mathbb{A})$. *Then there exists an automorphic representation* Π *of* $\mathrm{GL}_N(\mathbb{A})$ *for the appropriate* N *such that* Π_v *is the local Langlands lift of* π_v *for all archimedean places* v *and almost all non-archimedean places* v *where* π_v *is unramified.*

In these examples the local Langlands correspondence is not understood at the places v where π_v is ramified and so we must use the technique of multiplicativity and stability of the local γ-factors as outlined in Section 3. Multiplicativity has been established in generality by Shahidi [19] and in our first paper [2] we relied on the stability of γ-factors for SO_{2n+1} from [5]. Recently Shahidi has established an expression for his local coefficients as Mellin transforms of Bessel functions in some generality, and in particular in the cases at hand one can combine this with the results of [5] to obtain the necessary stability in the other cases, leading to the extension of the lifting to the other split classical groups [3].

2. *Tensor products.* Let $H = \mathrm{GL}_m \times \mathrm{GL}_n$. Then ${}^L\mathrm{H} = \mathrm{GL}_m(\mathbb{C}) \times \mathrm{GL}_n(\mathbb{C})$. Then there is a natural simple tensor product map from $\mathrm{GL}_m(\mathbb{C}) \times \mathrm{GL}_n(\mathbb{C})$ to $\mathrm{GL}_{mn}(\mathbb{C})$. The associated functoriality from $\mathrm{GL}_n \times \mathrm{GL}_m$ to GL_{mn} is the *tensor product lifting.* Now the associated local lifting is understood in principle since the local Langlands conjecture for GL_n has been solved. The question of global functoriality has been recently solved in the cases of $\mathrm{GL}_2 \times \mathrm{GL}_2$ to GL_4 by Ramakrishnan [17] and $\mathrm{GL}_2 \times \mathrm{GL}_3$ to GL_6 by Kim and Shahidi [15, 16].

Theorem. [17, 15] *Let* π_1 *be a cuspidal representation of* $\mathrm{GL}_2(\mathbb{A})$ *and* π_2 *a cuspidal representation of* $\mathrm{GL}_2(\mathbb{A})$ *(respectively* $\mathrm{GL}_3(\mathbb{A})$*). Then there is an automorphic representation* Π *of* $\mathrm{GL}_4(\mathbb{A})$ *(respectively* $\mathrm{GL}_6(\mathbb{A})$*) such that* Π_v *is the local tensor product lift of* $\pi_{1,v} \times \pi_{2,v}$ *at all places* v.

In both cases the authors are able to characterize when the lift is cuspidal.

In the case of Ramakrishnan [17] $\pi = \pi_1 \times \pi_2$ with each π_i cuspidal representation of $\mathrm{GL}_2(\mathbb{A})$ and Π is to be an automorphic representation of $\mathrm{GL}_4(\mathbb{A})$. To apply the Converse Theorem Ramakrishnan needs to control the analytic properties of $L(s, \Pi \times \pi')$ for π' cuspidal representations of $\mathrm{GL}_1(\mathbb{A})$ and $\mathrm{GL}_2(\mathbb{A})$, that is, the Rankin triple product L-functions $L(s, \Pi \times \pi') = L(s, \pi_1 \times \pi_2 \times \pi')$. This he was able to do using a combination of results on the integral representation for this L-function due to Garrett, Rallis and Piatetski-Shapiro, and Ikeda and the work of Shahidi on the Langlands-Shahidi method.

In the case of Kim and Shahidi [15, 16] π_2 is a cuspidal representation of $\mathrm{GL}_3(\mathbb{A})$. Since the lifted representation Π is to be an automorphic representation of $\mathrm{GL}_6(\mathbb{A})$, to apply the Converse Theorem they must control the analytic properties of $L(s, \Pi \times \pi') = L(s, \pi_1 \times \pi_2 \times \pi')$ where now π' must run over appropriate cuspidal representations of $\mathrm{GL}_m(\mathbb{A})$ with $m = 1, 2, 3, 4$. The control of these triple products is an application of the Langlands-Shahidi method of analysing L-functions and

involves coefficients of Eisenstein series on GL_5, Spin_{10}, and simply connected E_6 and E_7 [15, 21]. We should note that even though the complete local lifting theory is understood, they still use a highly ramified twist to control the global properties of the L-functions involved. They then show that their lifting is correct at all local places by using a base change argument.

3. *Symmetric powers.* Now take $H = \mathrm{GL}_2$, so ${}^L\mathrm{H} = \mathrm{GL}_2(\mathbb{C})$. For each $n \geq 1$ there is the natural symmetric n-th power map $sym^n : \mathrm{GL}_2(\mathbb{C}) \to \mathrm{GL}_{n+1}(\mathbb{C})$. The associated functoriality is the *symmetric power lifting* from representations of GL_2 to representations of GL_{n+1}. Once again the local symmetric powers liftings are understood in principle thanks to the solution of the local Langlands conjecture for GL_n. The global symmetric square lifting, so GL_2 to GL_3, is an old theorem of Gelbart and Jacquet. Recently, Kim and Shahidi have shown the existence of the global symmetric cube lifting from GL_2 to GL_4 [15] and then Kim followed with the global symmetric fourth power lifting from GL_2 to GL_5 [14].

Theorem. [15, 14] *Let π be a cuspidal automorphic representation of $\mathrm{GL}_2(\mathbb{A})$. Then there exists an automorphic representation Π of $\mathrm{GL}_4(\mathbb{A})$ (resp. $\mathrm{GL}_5(\mathbb{A})$) such that Π_v is the local symmetric cube (resp. symmetric fourth power) lifting of π_v.*

In either case, Kim and Shahidi have been able to give a very interesting characterization of when the image is in fact cuspidal [15, 16].

The original symmetric square lifting of Gelbart and Jacquet indeed used the converse theorem for GL_3. For Kim and Shahidi, the symmetric cube was deduced from the functorial $\mathrm{GL}_2 \times \mathrm{GL}_3$ tensor product lift above [15, 16] and did not require a new use of the Converse Theorem. For the symmetric fourth power lift, Kim first used the Converse Theorem to establish the *exterior square* lift from GL_4 to GL_6 by the method outlined above and then combined this with the symmetric cube lift to deduce the symmetric fourth power lift [14].

5. Applications

These new examples of functoriality have already had many applications. We will discuss the primary applications in parallel with our presentation of the examples. k remains a number field.

1. Classical groups: The applications so far of the lifting from classical groups to GL_n have been "internal" to the theory of automorphic forms. In the case of the lifting from SO_{2n+1} to GL_{2n}, once the weak lift is established, then the theory of Ginzburg, Rallis, and Soudry [8] allows one to show that this weak lift is indeed a strong lift in the sense that the local components Π_v at those $v \in S$ are completely determined and to completely characterize the image locally and globally. This will be true for the liftings from the other classical groups as well. Once one knows that these lifts are rigid, then one can begin to define and analyse the local lift for ramified representations by setting the lift of π_v to be the Π_v determined by the global lift. This is the content of the papers of Jiang and Soudry [12, 13] for the case of $\mathrm{H} = SO_{2n+1}$. In essence they show that this local lift satisfies the relations on L-functions that one expects from functoriality and then deduce the *local Langlands conjecture for* SO_{2n+1} from that for GL_{2n}. We refer to their papers for more detail

and precise statements.

2. Tensor product lifts: Ramakrishnan's original motivation for establishing the tensor product lifting from $\mathrm{GL}_2 \times \mathrm{GL}_2$ to GL_4 was to prove the multiplicity one conjecture for SL_2 of Langlands and Labesse.

Theorem. [17] *In the spectral decomposition*

$$L^2_{cusp}(\mathrm{SL}_2(k)\backslash \mathrm{SL}_2(\mathbb{A})) = \bigoplus m_\pi \pi$$

into irreducible cuspidal representations, the multiplicities m_π are at most one.

This was previously known to be true for GL_n and false for SL_n for $n \geq 3$. For further applications, for example to the Tate conjecture, see [17].

The primary application of the tensor product lifting from $\mathrm{GL}_2 \times \mathrm{GL}_3$ to GL_6 of Kim and Shahidi was in the establishment of the symmetric cube lifting and through this the symmetric fourth power lifting, so the applications of the symmetric power liftings outlined below are applications of this lifting as well.

3. Symmetric powers: It was early observed that the existence of the symmetric power liftings of GL_2 to GL_{n+1} for all n would imply the Ramanujan-Petersson and Selberg conjectures for modular forms. Every time a symmetric power lift is obtained we obtain better bounds towards Ramanujan. The result which follows from the symmetric third and fourth power lifts of Kim and Shahidi is the following.

Theorem. [16] *Let π be a cuspidal representation of $\mathrm{GL}_2(\mathbb{A})$ such that the symmetric cube lift of π is again cuspidal. Let $diag(\alpha_v, \beta_v)$ be the Satake parameter for an unramified local component. Then $|\alpha_v|, |\beta_v| < q_v^{1/9}$. If in addition the fourth symmetric power lift is not cuspidal, the full Ramanujan conjecture is valid.*

The corresponding statement at infinite places, i.e., the analogue of the Selberg conjecture on the eigenvalues of Mass forms, is also valid [14]. Estimates towards Ramanujan are a staple of improving any analytic number theoretic estimates obtained through spectral methods. Both the $1/9$ non-archimedean and $1/9$ archimedean estimate towards Ramanujan above were applied in obtaining the precise form of the exponent in our recent result with Sarnak breaking the convexity bound for twisted Hilbert modular L-series in the conductor aspect, which in turn was the key ingredient in our work on Hilbert's eleventh problem for ternary quadratic forms. Similar in spirit are the applications by Kim and Shahidi to the hyperbolic circle problem and to estimates on sums of shifted Fourier coefficients [15].

In addition Kim and Shahidi were able to obtain results towards the Sato-Tate conjecture.

Theorem. [16] *Let π be a cuspidal representation of $\mathrm{GL}_2(\mathbb{A})$ with trivial central character. Let $diag(\alpha_v, \beta_v)$ be the Satake parameter for an unramified local component and let $a_v = \alpha_v + \beta_v$. Assuming π satisfies the Ramanujan conjecture, there are sets $T^{\pm}$ of positive lower density for which $a_v > 2\cos(2\pi/11) - \epsilon$ for all $v \in T^+$ and $a_v < -2\cos(2\pi/11) + \epsilon$ for all $v \in T^-$. [Note: $2\cos(2\pi/11) = 1.68...$]*

Kim and Shahidi have other conditional applications of their liftings such as the conditional existence of Siegel modular cusp forms of weight 3 (assuming Arthur's multiplicity formula for Sp_4). We refer the reader to [15] for details on these applications and others.

References

[1] A. Borel, Automorphic L-functions. *Proc. Symp. Pure Math.* **33**, Part 2, (1979), 27–61.

[2] J.W. Cogdell, H. Kim, I.I. Piatetski-Shapiro, and F. Shahidi, On lifting from classical groups to GL_n. *Publ. Math. IHES* **93** (2001), 5–30.

[3] J.W. Cogdell, H. Kim, I.I. Piatetski-Shapiro, and F. Shahidi, On lifting from classical groups to GL_n, II, in preparation.

[4] J.W. Cogdell and I.I. Piatetski-Shapiro, Converse theorems for GL_n, I. *Publ. Math. IHES* **79** (1994), 157–214.

[5] J.W. Cogdell and I.I. Piatetski-Shapiro, Stability of gamma factors for SO_{2n+1}. *manuscripta math.* **95** (1998) 437–461.

[6] J.W. Cogdell and I.I. Piatetski-Shapiro, Converse theorems for GL_n, II. *J. reine angew. Math.* **507** (1999), 165–188.

[7] S. Gelbart and F. Shahidi, Boundedness of automorphic L-functions in vertical strips. *Journal of the AMS* **14** (2001), 79–107.

[8] D. Ginzburg, S. Rallis, and D. Soudry, Generic automorphic forms on SO_{2n+1}: functorial lift to GL_{2n}, endoscopy, and base change. *Internat. Math. Res. Notices.* no. **14**(2001), 729–764.

[9] H. Jacquet, I.I. Piatetski-Shapiro, and J. Shalika, Rankin-Selberg convolutions, *Amer. J. Math.*, **105** (1983), 367–464.

[10] H. Jacquet and J. Shalika, A lemma on highly ramified ε-factors, *Math. Ann.*, **271** (1985), 319–332.

[11] H. Jacquet and J. Shalika, Rankin-Selberg convolutions: Archimedean theory, in *Festschrift in Honor of I.I. Piatetski-Shapiro*, Part I, Weizmann Science Press, Jerusalem, 1990, 125–207.

[12] D. Jiang and D. Soudry, The local converse theorem for SO_{2n+1} and applications. *Ann. of Math*, to appear.

[13] D. Jiang and D. Soudry, Generic representations and local Langlands reciprocity law for p-adic SO_{2n+1}. *Preprint* (2001).

[14] H. Kim, Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2. *Preprint* (2000).

[15] H. Kim and F. Shahidi, Functorial products for $\mathrm{GL}_2 \times \mathrm{GL}_3$ and symmetric cube for GL_2. *Ann. of Math.*, to appear.

[16] H. Kim and F. Shahidi, Cuspidality of symmetric powers with applications. *Duke Math. J.*, **112** (2002), to appear.

[17] D. Ramakrishnan, Modularity of the Rankin-Selberg L-series, and multiplicity one for SL_2. *Ann. of Math. (2)* **152** (2000), 45–111.

[18] F. Shahidi, A proof of Langlands' conjecture on Plancherel measures; complementary series for p-adic groups. *Ann. of Math. (2)* **132** (1990), 273–330.

[19] F. Shahidi, On multiplicativity of local factors. *Festschrift in honor of I. I. Piatetski-Shapiro*, Part II (Ramat Aviv, 1989), 279–289. Israel Math. Conf. Proc., 3, Weizmann, Jerusalem, 1990.

[20] F. Shahidi, Local coefficients as Mellin transforms of Bessel functions; Towards a general stability. *Preprint* (2002).

[21] F. Shahidi, Automorphic L-Functions and Functoriality. These *Proceedings.*

ICM 2002 · Vol. II · 129–138

Constructing and Counting Number Fields

H. Cohen*

Abstract

In this paper we give a survey of recent methods for the asymptotic and exact enumeration of number fields with given Galois group of the Galois closure. In particular, the case of fields of degree up to 4 is now almost completely solved, both in theory and in practice. The same methods also allow construction of the corresponding complete tables of number fields with discriminant up to a given bound.

2000 Mathematics Subject Classification: 11R16, 11R29, 11R45, 11Y40.
Keywords and Phrases: Discriminants, Number field tables, Kummer theory.

1. Introduction

Let K be a number field considered as a fixed base field, $\overline{K}$ an algebraic closure of K, and G a transitive permutation group on n letters. We consider the set $\mathcal{F}_{K,n}(G)$ of all extensions L/K of degree n with $L \subset \overline{K}$ such that the Galois group of the Galois closure $\tilde{L}$ of L/K viewed as a permutation group on the set of embeddings of L into $\tilde{L}$ is permutation isomorphic to G (i.e, $n/m(G)$ times the number of extensions up to K-isomorphism, where $m(G)$ is the number of K-automorphisms of L). We write

$$N_{K,n}(G,X) = |\{L \in \mathcal{F}_n(G),\ |\mathcal{N}(\mathfrak{d}(L/K))| \le X\}|\,,$$

where $\mathfrak{d}(L/K)$ denotes the relative ideal discriminant and $\mathcal{N}$ the absolute norm. The aim of this paper is to give a survey of new methods, results, and conjectures on asymptotic and exact values of this quantity. It is usually easy to generalize the results to the case where the behavior of a *finite* number of places of K in the extension L/K is specified, for example if $K = \mathbb{Q}$ when the signature (R_1, R_2) of L is specified, with $R_1 + 2R_2 = n$.

Remarks.

*Laboratoire A2X, Institut de Mathématiques, Université Bordeaux I, 351 Cours de la Libération, 33405 TALENCE Cedex, France. E-mail: cohen@math.u-bordeaux.fr

1. It is often possible to give additional main terms and rather good error terms instead of asymptotic formulas. However, even in very simple cases such as $G = S_3$, this is not at all easy.
2. The methods which lead to exact values of $N_{K,n}(G, X)$ always lead to algorithms for computing the corresponding *tables*, evidently only when $N_{K,n}(G, X)$ is not too large in comparison to computer memory, see for example [8] and [10].

General conjectures on the subject have been made by several authors, for example in [3]. The most precise are due to G. Malle (see [24], [25]). We need the following definition.

Definition 1.1. *For any element $g \in S_n$ different from the identity, define the index* $\mathrm{ind}(g)$ *of g by the formula* $\mathrm{ind}(g) = n - |\text{orbits of } g|$. *We define the index $i(G)$ of a transitive subgroup G of S_n by the formula*

$$i(G) = \min_{g \in G,\ g \neq 1} \mathrm{ind}(g) \ .$$

Examples.

1. The index of a transposition is equal to 1, and this is the lowest possible index for a nonidentity element. Thus $i(S_n) = 1$.
2. If G is an Abelian group, and if ℓ is the smallest prime divisor of $|G|$, then $i(G) = |G|(1 - 1/\ell)$.

Conjecture 1.2. *For each number field K and transitive group G on n letters as above, there exist a strictly positive integer $b_K(G)$ and a strictly positive constant $c_K(G)$ such that*

$$N_{K,n}(G, X) \sim c_K(G)\, X^{1/i(G)} (\log X)^{b_K(G)-1} \ .$$

In [25], Malle gives a precise conjectural value for the constant $b_K(G)$ which is too complicated to be given here.

Remarks.

1. This conjecture is completely out of reach since it implies the truth of the inverse Galois problem for number fields.
2. If true, this conjecture implies that for any *composite* n, the proportion of S_n-extensions of K of degree n among all degree n extensions is strictly less than 1 (but strictly positive), contrary to the case of *polynomials.*

The following results give support to the conjecture (see [2], [9], [18], [19], [20], [21], [22], [23], [28], [30]).

Theorem 1.3. *We will say that the above conjecture is true in the weak sense if there exists $c_K(G) > 0$ such that for all $\varepsilon > 0$ we have*

$$c_K(G) \cdot X^{1/i(G)} < N_{K,n}(G, X) < X^{1/i(G)+\varepsilon} \ .$$

1. *(Mäki, Wright). The conjecture is true for all Abelian groups G.*
2. *(Davenport-Heilbronn, Datskovsky-Wright). The conjecture is true for $n = 3$ and $G = S_3$.*

3. *(Cohen-Diaz-Olivier). The conjecture is true for* $n = 4$ *and* $G = D_4$.
4. *(Bhargava, Yukie). The conjecture is true for* $n = 4$ *and* $G = S_4$, *in the weak sense if* $K \neq \mathbb{Q}$.
5. *(Klüners-Malle). The conjecture is true in the weak sense for all nilpotent groups.*
6. *(Kable-Yukie). The conjecture is true in the weak sense for* $n = 5$ *and* $G = S_5$.

The methods used to prove these results are quite diverse. In the case of Abelian groups G, one could think that class field theory gives all the answers so nothing much would need to be done. This is not at all the case, and in fact Kummer theory is usually more useful. In addition, Kummer theory allows us more generally to study solvable groups. We will look at this method in detail.

Apart from Kummer theory and class field theory, the other methods have a different origin and come from the classification of *orders* of degree n, interpreted through suitable classes of *forms*. This can be done at a very clever but still elementary level when the base field is $\mathbb{Q}$, and includes the remarkable achievement of M. Bhargava in 2001 for quartic orders. Over arbitrary K, one needs to use and develop the theory of prehomogeneous vector spaces, initiated at the end of the 1960's by Sato and Shintani (see for example [26] and [27]), and used since with great success by Datskovsky-Wright, and more recently by Wright-Yukie (see [29]), Yukie and Kable-Yukie.

2. Kummer theory

This method applies only to Abelian, or more generally solvable extensions.

2.1. Why not class field theory?

It is first important to explain why class field theory, which is supposed to be a complete theory of Abelian extensions, does not give an answer to counting questions. Let us take the very simplest example of quadratic extensions, thus with $G = C_2$. A trivial class-field theoretic argument gives the exact formula

$$N_{K,2}(C_2, X) = -1 + \sum_{\mathcal{N}(\mathfrak{a}) \leq X} 2^{\mathrm{rk}(Cl_{\mathfrak{a}}^+(K))} M_K\left(\frac{X}{\mathcal{N}(\mathfrak{a})}\right) ,$$

where $\mathfrak{a}$ runs over all integral ideals of K of norm less than or equal to X, $Cl_{\mathfrak{a}}^+(K)$ denotes the narrow ray class group modulo $\mathfrak{a}$, $\mathrm{rk}(G)$ denotes the 2-rank of an Abelian group G, and $M_K(n)$ is the generalization to number fields of the summatory function $M(n)$ of the Möbius function.

This formula is completely explicit, the quantities $Cl_{\mathfrak{a}}^+(K)$ and the function $M_K(n)$ are algorithmically computable with reasonable efficiency, so we can compute $N_{K,2}(C_2, X)$ for reasonably small values of X in this way. Unfortunately, this formula has two important drawbacks.

The first and essential one is that, if we want to deduce from it asymptotic information on $N_{K,2}(C_2, X)$, we need to control $\mathrm{rk}(Cl_{\mathfrak{a}}^+(K))$, which can be done,

although rather painfully, but we also need to control $M_K(n)$, which *cannot* be done (recall for instance that the Riemann Hypothesis can be formulated in terms of this function).

The second drawback is that, even for exact computation it is rather inefficient, compared to the formula that we obtain from Kummer theory. Thus, even though Kummer theory is used in a crucial way for the constructions needed in the proofs of class field theory, it must not be discarded once this is done since the formula that it gives are much more useful, at least in our context.

2.2. Quadratic extensions

As an example, let us see how to treat quadratic extensions using Kummer theory instead of class field theory. Of course in this case Kummer theory is trivial since it tells us that quadratic extensions of K are parameterized by K^*/K^{*2} minus the unit class. This is not explicit enough. By writing for any $\alpha \in K^*$, $\alpha\mathbb{Z}_K = \mathfrak{a}\mathfrak{q}^2$ with $\mathfrak{a}$ an integral squarefree ideal, it is clear that K^*/K^{*2} is in one-to-one correspondence with pairs $(\mathfrak{a}, \overline{u})$, where $\mathfrak{a}$ are integral squarefree ideals whose ideal class is a square, and $\overline{u}$ is an element of the so-called Selmer group of K, i.e., the group of elements $u \in K^*$ such that $u\mathbb{Z}_K = \mathfrak{q}^2$ for some ideal $\mathfrak{q}$, divided by K^{*2}. We can then introduce the Dirichlet series $\Phi_{K,2}(C_2, s) = \sum_L \mathcal{N}(\mathfrak{d}(L/K))^{-s}$, where the sum is over quadratic extensions L/K in $\overline{K}$. A number of not completely trivial combinatorial and number-theoretic computations (see [9]) lead to the explicit formula

$$\Phi_{K,2}(C_2, s) = -1 + \frac{2^{-r_2}}{\zeta_K(2s)} \sum_{\mathfrak{c}|2} \frac{\mathcal{N}(2/\mathfrak{c})}{\mathcal{N}(2/\mathfrak{c})^s} \sum_{\chi} L_K(\chi, s) \ ,$$

where χ runs over all quadratic characters of the ray class group $Cl_{\mathfrak{c}^2}(K)$ and $L_K(\chi, s)$ is the ordinary Dirichlet-Hecke L-function attached to χ.

There are two crucial things to note in this formula. First of all, the sum on $\mathfrak{c}$ is only on integral ideals dividing 2, so is finite and very small. Thus, $\Phi_{K,2}(C_2, s)$ is a finite linear combination of Euler products, and can directly be used much more efficiently than the formula coming from class field theory to compute $N_{K,2}(C_2, X)$ exactly. For example (but this of course does not need the above machinery) we obtain $N_{\mathbb{Q},2}(C_2, 10^{25}) = 607927101854026628651779 5$.

Second, since $L_K(\chi, s)$ extends to a meromorphic function in the whole complex plane with no pole if χ is not a trivial character, the polar part of $\Phi_{K,2}(C_2, s)$, which is the only thing that we need for an asymptotic formula, comes only from the contributions of the trivial characters, in which case $L_K(\chi, s)$ is equal to $\zeta_K(s)$ times a finite number of Euler factors. We thus obtain

$$N_{K,2}(C_2, X) \sim \frac{1}{2^{r_2}} \frac{\zeta_K(1)}{\zeta_K(2)} X \ ,$$

where $\zeta_K(1)$ is a convenient abuse of notation for the residue of $\zeta_K(s)$ at $s = 1$. Apparently this simple result was first obtained by Datskovsky-Wright in [18], although their proof is different.

2.3. General finite Abelian extensions

The same method can in principle be applied to any finite Abelian group G. I say "in principle", because in practice several problems arise. For the base field $K = \mathbb{Q}$, a complete and explicit solution was given by Mäki in [23]. For a general base field, a solution has been given by Wright in [28], but the problem with his solution is that the constant $c_K(G)$, although given as a product of local contributions, cannot be computed explicitly without a considerable amount of additional work. It is always a finite linear combination of Euler products.

In joint work with F. Diaz y Diaz and M. Olivier, using Kummer theory in a manner analogous but much more sophisticated than the case of quadratic extensions, we have computed completely explicitly the constants $c_K(G)$ for $G = C_\ell$ the cyclic group of prime order ℓ, for $G = C_4$ and for $G = V_4 = C_2 \times C_2$. Although our papers are perhaps slightly too discursive, to give an idea the total number of pages for these three results exceeds 100. We refer to [7], [13], [11], [15], [16] for the detailed proofs, and to [12] and [14] for surveys and tables of results. We mention here the simplest one, for $G = V_4$. We have

$$N_{K,4}(V_4, X) \sim c_K(V_4)\, X^{1/2} \log^2 X \quad \text{with}$$

$$c_K(V_4) = \frac{1}{48 \cdot 4^{r_2}} \zeta_K(1)^3 \prod_{\mathfrak{p}} \left(1 + \frac{3}{\mathcal{N}\mathfrak{p}}\right) \left(1 - \frac{1}{\mathcal{N}\mathfrak{p}}\right)^3 \prod_{\mathfrak{p} \mid 2\mathbb{Z}_K} \frac{1 + \dfrac{4}{\mathcal{N}\mathfrak{p}} + \dfrac{2}{\mathcal{N}\mathfrak{p}^2} + \dfrac{1}{\mathcal{N}\mathfrak{p}^3} - \dfrac{(1 - 1/\mathcal{N}\mathfrak{p}^2)e(\mathfrak{p}) + (1 + 1/\mathcal{N}\mathfrak{p})^2}{\mathcal{N}\mathfrak{p}^{e(\mathfrak{p})+1}}}{1 + \dfrac{3}{\mathcal{N}\mathfrak{p}}}.$$

Of course, the main difficulty is to compute correctly the local factor at 2.

As usual, we can use our methods to compute very efficiently the N function. For example, we obtain (see [4]):

$$N_{\mathbb{Q},3}(C_3, 10^{37}) = 501310370031289126,$$
$$N_{\mathbb{Q},4}(C_4, 10^{32}) = 1220521363354404,$$
$$N_{\mathbb{Q},4}(V_4, 10^{36}) = 22956815681347605884.$$

2.4. Dihedral D_4-extensions

We can also apply our method to solvable extensions. The case of quartic D_4-extensions, where D_4 is the dihedral group of order 8, is especially simple and pretty. Such an extension is imprimitive, i.e., is a quadratic extension of a quadratic extension. Conversely, imprimitive quartic extensions are either D_4-extensions, or Abelian with Galois group C_4 or V_4. These can easily be counted as explained above, and in any case will not contribute to the main term of the asymptotic formula, so they can be neglected (or subtracted for exact computations). Since we have treated completely the case of quadratic extensions, it is just a matter of showing

that we are allowed to sum over quadratic extensions of the base field to obtain the desired asymptotic formula (for the exact formula nothing needs to be proved), and this is not difficult. In this way, we obtain that $N_{K,4}(D_4, X) \sim c_K(D_4)\,X$ for an explicit constant $c_K(D_4)$ (in fact we obtain an error term $O(X^{3/4+\varepsilon})$). This result is new even for $K = \mathbb{Q}$, although its proof not very difficult. In the case $K = \mathbb{Q}$, we have for instance

$$c_{\mathbb{Q}}(D_4) = \frac{6}{\pi^2} \sum_D \frac{2^{-r_2(D)}}{D^2} \frac{L(\left(\frac{D}{\cdot}\right), 1)}{L(\left(\frac{D}{\cdot}\right), 2)} = 0.1046520224\ldots\,,$$

where the sum is over fundamental discriminants D, $r_2(D) = r_2(\mathbb{Q}(\sqrt{D}))$, and $L(\left(\frac{D}{\cdot}\right), s)$ is the usual Dirichlet series for the character $\left(\frac{D}{\cdot}\right)$.

Remark. In the Abelian case, it is possible to compute the Euler products which occur to hundreds of decimal places if desired using almost standard zeta-product expansions, see for example [6]. Unfortunately, we do not know if it is possible to express $c_{\mathbb{Q}}(D_4)$ as a finite linear combination of Euler products (or at least as a rapidly convergent infinite series of such), hence we have only been able to compute 9 or 10 decimal places of this constant. We do not see any practical way of computing 20 decimals, say.

Our method also allows us to compute $N_{\mathbb{Q},4}(D_4, X)$ exactly. However, here a miracle occurs: when k is a quadratic field, in the formula that we have given above for $\Phi_{k,2}(C_2, s)$ all the quadratic characters χ which we need are *genus characters* in the sense of Gauss, in other words there is a decomposition

$$L_k(\chi, s) = L(\left(\tfrac{d_1}{\cdot}\right), s) L(\left(\tfrac{d_2}{\cdot}\right), s)$$

into a product of two suitable ordinary Dirichlet L-series. This gives a very fast method for computing $N_{\mathbb{Q},4}(D_4, X)$, and in particular we have been able to compute $N_{\mathbb{Q},4}(D_4, 10^{17}) = 10465196820067560$.

We can also count the number of extensions with a given signature. The method is completely similar, but here not all characters are genus characters. In fact, it is only necessary to add a single nongenus character to obtain all the necessary ones, but everything is completely explicit, and closely related to the *rational quartic reciprocity law.* I refer to [5] for details.

2.5. Other solvable extensions

We can also prove some partial results in the case where $G = A_4$ or $G = S_4$ (of course the results for S_4 are superseded by Bhargava's for $K = \mathbb{Q}$, and by Yukie's for general K; still, the method is also useful for exact computations), see [17].

In the case of quartic A_4 and S_4-extensions (or, for that matter, of cubic S_3-extensions), we use the diagram involving the cubic resolvent (the quadratic one for S_3-extensions), also called the Hasse diagram. We then have a situation which bears some analogies with the D_4 case. The differences are as follows. Instead of having to sum over quadratic extensions of the base field K, we must sum over cubic extensions, cyclic for A_4 and noncyclic for S_4. As in the D_4-case, we then have to consider quadratic extensions of these cubic fields, but generated by an element of

square norm. It is possible to go through the exact combinatorial and arithmetic computation of the corresponding Dirichlet series, the cubic field being fixed. This in particular uses some amusing local class field theory. As in the D_4 case, we then obtain the Dirichlet generating series for discriminants of A_4 (resp., S_4) extensions by summing the series over the corresponding cubic fields.

Unfortunately, we cannot obtain from this any asymptotic formula. The reason is different in the A_4 and the S_4 case. In the A_4 case, the rightmost singularity of the Dirichlet series is at $s = 1/2$. Unfortunately, this is simultaneously the main singularity of each individual Dirichlet series, and also that of the generating series for cyclic cubic fields. Thus, although the latter is well understood, it seems difficult (but not totally out of reach) to paste things together. On the other hand, we can do two things rigorously in this case. First, we can prove an asymptotic formula for A_4-extensions having a *fixed* cubic resolvent. Tables show that the formula is very accurate. Second, we can use our formula to compute $N_{K,4}(S_4, X)$ exactly. For instance, we have computed $N_{\mathbb{Q},4}(A_4, 10^{16}) = 218369252$. This computation is much slower than in the D_4-case, because we do not have the miracle of genus characters, and we must compute the class and unit group of all the cyclic cubic fields.

In the S_4 case, the situation is different. The main singularity of each individual Dirichlet series is still at $s = 1/2$ (because of the square norm condition), and the rightmost singularity of the generating series for noncyclic cubic fields is at $s = 1$, so the situation looks better (and analogous to the D_4 situation with s replaced by $s/2$). Unfortunately, as already mentioned we know almost nothing about the generating series for noncyclic cubic fields, a fortiori with coefficients. So we cannot go further in the asymptotic analysis. As in the A_4 case, however, we can compute exactly either the number of S_4-extensions corresponding to a fixed cubic resolvent, or even $N_{K,4}(S_4, X)$ itself. The problem is that here we must compute class and unit groups of all noncyclic cubic fields of discriminant up to X, while cyclic cubic fields of discriminant up to X are much rarer, of the order of $X^{1/2}$ instead. We have thus not been able to go very far and obtained for example $N_{\mathbb{Q},4}(S_4, 10^7) = 6541232$.

3. Prehomogeneous vector spaces

The other methods for studying $N_{K,n}(G, X)$ are two closely related methods: one is the use of generalizations of the Delone-Fadeev map, which applies when $K = \mathbb{Q}$. The other, which can be considered as a generalization of the first, is the use of the theory of prehomogeneous vector spaces, initiated by Sato and Shintani in the 1960's.

3.1. Orders of small degree

We briefly give a sketch of the first method. We would first like to classify *quadratic* orders. It is well known that, through their discriminant, such orders are in one-to-one correspondence with the subset of nonsquare elements of $\mathbb{Z}$ congruent

to 0 or 1 modulo 4, on which $\mathrm{SL}_1(\mathbb{Z})$ (the trivial group) acts. Thus, for fixed discriminant, the orbits are finite (in fact of cardinality 0 or 1). For *maximal* orders, we need to add local arithmetic conditions at each prime p, which are easy for $p > 2$, and slightly more complicated for $p = 2$.

We do the same for small higher degrees. For *cubic* orders, the classification is due to Davenport-Heilbronn (see [19], [20]). These orders are in one-to-one correspondence with a certain subset of $\mathrm{Sym}^3(\mathbb{Z}^2)$, i.e., binary cubic forms, on which $\mathrm{SL}_2(\mathbb{Z})$ acts. Since once again the difference in "dimensions" is $4 - 3 = 1$, for fixed discriminant the orbits are finite, at least generically. For maximal orders, we again need to add local arithmetic conditions at each prime p. These are easy to obtain for $p > 3$, but are a little more complicated for $p = 2$ and $p = 3$. An alternate way of explaining this is to say that a cubic order can be given by a *nonmonic* cubic equation, which is almost canonical if representatives are suitably chosen.

For *quartic* orders, the classification is due to M. Bhargava in 2001, who showed in complete detail how to generalize the above. These orders are now in one-to-one correspondence with a certain subset of $\mathbb{Z}^2 \otimes \mathrm{Sym}^2(\mathbb{Z}^3)$, i.e., pairs of ternary quadratic forms, on which $\mathrm{SL}_2(\mathbb{Z}) \times \mathrm{SL}_3(\mathbb{Z})$ acts. Once again the difference in "dimensions" is $2 \times 6 - (3 + 8) = 1$, so for fixed discriminant the orbits are finite, at least generically. For maximal orders, we again need to add local arithmetic conditions at each prime p, which Bhargava finds after some computation. An alternate way of explaining this is to say that a quartic order can be given by the intersection of two conics in the projective plane, the pencil of conics being almost canonical if representatives are suitably chosen.

For *quintic* orders, only part of the work has been done, by Bhargava and Kable-Yukie in 2002. These are in one-to-one correspondence with a certain subset of $\mathbb{Z}^4 \otimes \Lambda^2(\mathbb{Z}^5)$, i.e., quadruples of alternating forms in 5 variables, on which $\mathrm{SL}_4(\mathbb{Z}) \times \mathrm{SL}_5(\mathbb{Z})$ acts. Once again the difference in "dimensions" is $4 \times 10 - (15 + 24) = 1$, so for fixed discriminant the orbits are finite, at least generically. The computation of the local arithmetic conditions, as well as the justification for the process of point counting near the cusps of the fundamental domain has however not yet been completed.

Since prehomogeneous vector spaces have been completely classified, this theory does not seem to be able to apply to higher degree orders, at least directly.

3.2. Results

Using the above methods, and generalizations to arbitrary base fields, the following results have been obtained on the function $N_{K,n}(G, X)$ (many other deep and important results have also been obtained, but we fix our attention to this function). It is important to note that they seem out of reach using more classical methods such as Kummer theory or class field theory mentioned earlier.

Theorem 3.1. *Let K be a number field of signature (r_1, r_2), and as above write $\zeta_K(1)$ for the residue of the Dedekind zeta function of K at $s = 1$.*

1. (Davenport-Heilbronn [19], [20]). We have $N_{\mathbb{Q},3}(S_3, X) \sim c_{\mathbb{Q}}(S_3)\, X$ with

$$c_{\mathbb{Q}}(S_3) = \frac{1}{\zeta(3)} .$$

2. *(Datskovsky-Wright [18]). We have* $N_{K,3}(S_3, X) \sim c_K(S_3)\, X$ *with*

$$c_K(S_3) = \left(\frac{2}{3}\right)^{r_1-1} \left(\frac{1}{6}\right)^{r_2} \frac{\zeta_K(1)}{\zeta_K(3)} .$$

3. *(Bhargava [1], [2]). We have* $N_{\mathbb{Q},4}(S_4, X) \sim c_{\mathbb{Q}}(S_4)\, X$ *with*

$$c_{\mathbb{Q}}(S_4) = \frac{5}{6} \prod_p \left(1 + \frac{1}{p^2} - \frac{1}{p^3} - \frac{1}{p^4}\right) .$$

4. *(Yukie [30]). There exist two strictly positive constants* $c_1(K)$ *and* $c_2(K)$ *such that*

$$c_1\, X < N_{K,4}(S_4, X) < c_2\, X \log^2(X) .$$

Under some very plausible convergence assumptions we should have in fact $N_{K,4}(S_4, X) \sim c_K(S_4)\, X$ *with*

$$c_K(S_4) = 2 \left(\frac{5}{12}\right)^{r_1} \left(\frac{1}{24}\right)^{r_2} \prod_{\mathfrak{p}} \left(1 + \frac{1}{\mathcal{N}\mathfrak{p}^2} - \frac{1}{\mathcal{N}\mathfrak{p}^3} - \frac{1}{\mathcal{N}\mathfrak{p}^4}\right) .$$

5. *(Kable-Yukie [21]). There exists a strictly positive constant* c_1 *such that for all* $\varepsilon > 0$ *we have*

$$c_1\, X < N_{\mathbb{Q},5}(S_5, X) < X^{1+\varepsilon} .$$

Remark. It should also be emphasized that, although the above methods give important and deep results on $N_{K,n}(G, X)$ for certain groups G, they shed almost no light on the possible analytic continuation of the corresponding Dirichlet series of which $N_{K,n}(G, X)$ is the counting function. For example, in the simplest case where $K = \mathbb{Q}$, $n = 3$, and $G = S_3$, for which the result dates back to Davenport-Heilbronn, no one knows how to give an analytic continuation of the Dirichlet series $\sum_L |d(L)|^{-s}$ even to $\Re(s) = 1$ (the sum being over cubic fields in $\overline{\mathbb{Q}}$ and $d(L)$ being the absolute discriminant of L).

References

[1] M. Bhargava, *Higher composition laws*, PhD Thesis, Princeton Univ., June 2001.

[2] M. Bhargava, *Higher composition laws*, Proceedings ANTS V Conference, Sydney (2002), Lecture Notes in Comp. Sci., to appear.

[3] H. Cohen, *Advanced topics in computational number theory*, GTM **193**, Springer-Verlag, 2000.

[4] H. Cohen, *Comptage exact de discriminants d'extensions abéliennes*, J. Th. Nombres Bordeaux **12** (2000), 379–397.

[5] H. Cohen, *Enumerating quartic dihedral extensions of* $\mathbb{Q}$ *with signatures*, 32p., submitted.

[6] H. Cohen, *High precision computation of Hardy-Littlewood constants*, preprint available on the author's web page.

[7] H. Cohen, F. Diaz y Diaz and M. Olivier, *Densité des discriminants des extensions cycliques de degré premier*, C. R. Acad. Sci. Paris **330** (2000), 61–66.
[8] H. Cohen, F. Diaz y Diaz and M. Olivier, *Construction of tables of quartic fields using Kummer theory*, Proceedings ANTS IV, Leiden (2000), Lecture Notes in Computer Science **1838**, Springer-Verlag, 257–268.
[9] H. Cohen, F. Diaz y Diaz and M. Olivier, *Enumerating quartic dihedral extensions*, Compositio Math., 28p., to appear.
[10] H. Cohen, F. Diaz y Diaz and M. Olivier, *Constructing complete tables of quartic fields using Kummer theory*, Math. Comp., 11p., to appear.
[11] H. Cohen, F. Diaz y Diaz and M. Olivier, *On the density of discriminants of cyclic extensions of prime degree*, J. reine und angew. Math., 40p., to appear.
[12] H. Cohen, F. Diaz y Diaz and M. Olivier, *A Survey of Discriminant Counting*, Proceedings ANTS V Conference, Sydney (2002), Lecture Notes in Comp. Sci., 15p., to appear.
[13] H. Cohen, F. Diaz y Diaz and M. Olivier, *Cyclotomic extensions of number fields*, 14p., submitted.
[14] H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting discriminants of number fields*, 36p., submitted.
[15] H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting cyclic quartic extensions of a number field*, 30p., submitted.
[16] H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting biquadratic extensions of a number field*, 17p., submitted.
[17] H. Cohen, F. Diaz y Diaz and M. Olivier, *Counting A_4 and S_4 extensions of number fields*, 20p., in preparation.
[18] B. Datskovsky and D. J. Wright, *Density of discriminants of cubic extensions*, J. reine und angew. Math. **386** (1988), 116–138.
[19] H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields I*, Bull. London Math. Soc. **1** (1969), 345–348.
[20] H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields II*, Proc. Royal. Soc. A **322** (1971), 405–420.
[21] A. Kable and A. Yukie, *On the number of quintic fields*, preprint.
[22] J. Klüners and G. Malle, *Counting Nilpotent Galois Extensions*, submitted.
[23] S. Mäki, *On the density of abelian number fields*, Thesis, Helsinki, 1985.
[24] G. Malle, *On the distribution of Galois groups*, J. Number Theory, to appear.
[25] G. Malle, personal communication.
[26] T. Shintani, *On Dirichlet series whose coefficients are class numbers of integral binary cubic forms*, J. Math. Soc. Japan **24** (1972), 132–188.
[27] T. Shintani, *On zeta-functions associated with the vector space of quadratic forms*, J. Fac. Sci. Univ. Tokyo, Sec. 1a **22** (1975), 25–66.
[28] D. J. Wright, *Distribution of discriminants of Abelian extensions*, Proc. London Math. Soc. (3) **58** (1989), 17–50.
[29] D. J. Wright and A. Yukie, *Prehomogeneous vector spaces and field extensions*, Invent. Math. **110** (1992), 283–314.
[30] A. Yukie, *Density theorems related to prehomogeneous vector spaces*, preprint.

ICM 2002 · Vol. II · 139–148

Analyse p-adique et Représentations Galoisiennes

Jean-Marc Fontaine*

Abstract

The notion of a p-adic de Rham representation of the absolute Galois group of a p-adic field was introduced about twenty years ago (see e.g. [Fo93]). Three important results for this theory have been obtained recently: The structure theorem for the *almost C_p-representations*, the theorem *weakly admissible implies admissible* and the theorem *de Rham implies potentially semi-stable*. The proofs of the first two theorems are closely related to the study of a new kind of analytic groups, the *Banach-Colmez spaces* and the proof of the third uses deep results on *p-adic differential equations on the Robba ring*.

2000 Mathematics Subject Classification: 11F80, 11S25, 12H25, 14G22.
Keywords and Phrases: Galois representations, de Rham representations, Semi-stable representations, p-adic Banach spaces, p-adic differential equations.

1. Représentations p-adiques

1.1. – Dans tout ce qui suit, K est un corps de caractéristique 0, complet pour une valuation discrète, à corps résiduel k parfait de caractéristique $p > 0$. On choisit une clôture algébrique $\overline{K}$ de K, on note C son complété et $|\ |_p$ la valeur absolue de C normalisée par $|p|_p = p^{-1}$. On pose $G_K = \mathrm{Gal}(\overline{K}/K)$.

Une représentation banachique (de G_K) est un espace de Banach p-adique muni d'une action linéaire et continue de G_K. Avec comme morphismes les applications $\mathbb{Q}_p$-linéaires continues G_K-équivariantes, ces représentations forment une catégorie additive $\mathbb{Q}_p$-linéaire $\mathcal{B}(G_K)$.

Une *C-représentation (de G_K)* est un C-espace vectoriel de dimension finie muni d'une action semi-linéaire et continue de G_K. Lorsque k est fini, la catégorie $\mathrm{Rep}_C(G_K)$ des C-représentations s'identifie à une sous-catégorie pleine de $\mathcal{B}(G_K)$:

* Institut Universitaire de France et UMR 8628 du CNRS, Mathématique, Université de Paris-Sud, Bâtiment 425, 91405 ORSAY Cedex, France. E-mail: fontaine@math.u-psud.fr

PROPOSITION [Fo00]. — *Supposons k fini. Si W_1 et W_2 sont des C-représentations, toute application $\mathbb{Q}_p$-linéaire continue G_K-équivariante de W_1 dans W_2 est C-linéaire.*

Disons que deux représentations banachiques S_1 et S_2 sont *presque isomorphes* s'il existe un triplet (V_1, V_2, α) où V_i est un sous-$\mathbb{Q}_p$-espace vectoriel de dimension finie de S_i, stable par G_K, et où $\alpha : S_1/V_1 \to S_2/V_2$ est un isomorphisme (dans $\mathcal{B}(G_K)$). Une *presque-C-représentation (de G_K)* est une représentation banachique qui est presque isomorphe à une C-représentation. On note $\mathcal{C}(G_K)$ la sous-catégorie pleine de $\mathcal{B}(G_K)$ dont les objets sont les presque-C-représentations. Elle contient la catégorie $\mathrm{Rep}_C(G_K)$ et la catégorie $\mathrm{Rep}_{\mathbb{Q}_p}(G_K)$ des *représentations p-adiques de dimension finie (de G_K)* comme sous-catégories pleines.

THÉORÈME A [Fo02]. — *Supposons k fini. La catégorie $\mathcal{C}(G_K)$ est abélienne. Il existe sur les objets de $\mathcal{C}(G_K)$ une unique fonction additive $dh : S \mapsto (d(S), h(S)) \in \mathbb{N} \times \mathbb{Z}$ telle que $dh(W) = (\dim_C W, 0)$ si W est une C-représentation et $dh(V) = (0, \dim_{\mathbb{Q}_p} V)$ si V est de dimension finie sur $\mathbb{Q}_p$.*

Si S et T sont des objets de $\mathcal{C}(G_K)$, les $\mathbb{Q}_p$-espaces vectoriels $\mathrm{Ext}^i_{\mathcal{C}(G_K)}(S,T)$ sont de dimension finie et sont nuls pour $i \notin \{0,1,2\}$. On a

$$\textstyle\sum_{i=0}^{2}(-1)^i \dim_{\mathbb{Q}_p} \mathrm{Ext}^i_{\mathcal{C}(G_K)}(S,T) = -[K : \mathbb{Q}_p]h(S)h(T).$$

1.2. — Soit [(1)] FR l'ensemble des suites $(x^{(n)})_{n\in\mathbb{N}}$ d'éléments de C vérifiant $(x^{(n+1)})^p = x^{(n)}$ pour tout n. Avec les lois

$$(x+y)^{(n)} = \lim_{m\mapsto\infty}(x^{(n+m)} + y^{(n+m)})^{p^m} \text{ et } (xy)^{(n)} = x^{(n)}y^{(n)}$$

c'est un corps algébriquement clos de caractéristique p, complet pour la valeur absolue définie par $|x| = |x^{(0)}|_p$ et on note R l'anneau de la valuation. Son corps résiduel s'identifie au corps résiduel $\overline{k}$ de $\overline{K}$. L'anneau $W(R)$ des vecteurs de Witt à coefficients dans R est intègre. Choisissons $\varepsilon, \pi \in R$ vérifiant $\varepsilon^{(0)} = 1$, $\varepsilon^{(1)} \neq 1$ et $\pi^{(0)} = p$ et, pour tout $a \in R$ notons $[a] = (a, 0, 0, \ldots)$ son représentant de Teichmüller dans $W(R)$. L'application $\theta : W(R) \to \mathcal{O}_C$ qui envoie $(a_0, a_1, \ldots)$ sur $\sum_{n\in\mathbb{N}} p^n a_n^{(n)}$ est un homomorphisme d'anneaux surjectif dont le noyau est l'idéal principal engendré par $\xi = [\pi] - p$. On note encore $\theta : W(R)[1/p] \to C$ l'application déduite en rendant p inversible. Rappelons que $B^+_{dR} = \varprojlim_{n\in\mathbb{N}} W(R)[1/p]/(\xi^n)$ et que le corps B_{dR} des périodes p-adiques est le corps des fractions de B^+_{dR}. Toute unité a de R s'écrit d'une manière unique sous la forme $a = a_0 a^+$ avec $a_0 \in \overline{k}$ et $|a^+ - 1| < 1$, la série $\sum_{n=1}^{+\infty}(-1)^{n+1}([a^+] - 1)^n/n$ converge dans B^+_{dR} vers un élément noté $\log[a]$; on pose $t = \log[\varepsilon]$. On a $B_{dR} = B^+_{dR}[1/t]$. On note A_{cris} le séparé complété pour la topologie p-adique de la sous-$W(R)$-algèbre de $W(R)[1/p]$ engendrée par les $\xi^m/m!$ pour $m \in \mathbb{N}$. Alors A_{cris} s'identifie à un sous-anneau de B^+_{dR} contenant t et on pose $B^+_{cris} = A_{cris}[1/p]$ et $B_{cris} = B^+_{cris}[1/t] \subset B_{dR}$. La série $\sum_{n=1}^{+\infty}(-1)^{n+1}\xi^n/np^n$ converge dans B^+_{dR} vers un élément $\log[\pi] = \log([\pi]/p)$ et on

(1) Voir [Fo88a] (resp. [Fo88b]) pour plus de détails sur la construction de B_{dR}, B_{cris} et B_{st} (resp. sur les représentations p-adiques de de Rham et potentiellement semi-stables).

note B_{st} la sous-B_{cris}-algèbre de B_{dR} engendrée par $\log[\pi]$. Pour tout $b \in R$ non nul, il existe $r, s \in \mathbb{Z}$, avec $s \geq 1$ et une unité a de R tels que $b^s = \pi^r a$ et on pose $\log[b] = (r\log[\pi] + \log[a])/s$. On a $B_{st} = B_{cris}[\log[b]]$ dès que b n'est pas une unité.

Soit $\mathcal{L}$ l'ensemble des extensions finies de K contenues dans $\overline{K}$. Pour tout $L \in \mathcal{L}$, on pose $G_L = \mathrm{Gal}(\overline{K}/L)$ et on note L_0 le corps des fractions de l'anneau des vecteurs de Witt à coefficients dans le corps résiduel de L. Le corps $\overline{K}$ se plonge de façon naturelle dans B_{dR}^+ et l'action de G_K s'étend de façon naturelle à B_{dR}, l'anneau B_{st} est stable par G_K. Pour tout $L \in \mathcal{L}$, on a $(B_{dR})^{G_L} = L$ tandis que $(B_{st})^{G_L} = L_0$ et l'application naturelle $L \otimes_{L_0} B_{st} \to B_{dR}$ est injective.

Pour toute représentation p-adique V de G_K de dimension finie h sur $\mathbb{Q}_p$, on pose $D_{dR}(V) = (\overline{K} \otimes_{\mathbb{Q}_p} V)^{G_K}$, $D_{st}(V) = (B_{st} \otimes_{\mathbb{Q}_p} V)^{G_K}$ et, si L est une extension finie de K contenue dans $\overline{K}$, $D_{st,L}(V) = (B_{st} \otimes_{\mathbb{Q}_p} V)^{G_L}$. On a $\dim_K D_{dR}(V) \leq h$ et on dit que V *est de de Rham* si on a l'égalité. C'est le cas si $\dim_{K_0} D_{st}(V) = h$ auquel cas on dit que V *est semi-stable.* C'est aussi le cas s'il existe $L \in \mathcal{L}$ tel que $\dim_{L_0} D_{st,L}(V) = h$, auquel cas on dit que V *est potentiellement semi-stable*, ou si l'on veut préciser L, que V *est L-semi-stable.*

THÉORÈME B. — *Toute représentation p-adique de G_K qui est de de Rham est potentiellement semi-stable.*

Soit $\overline{K}B_{st}$ le plus petit sous-anneau de B_{dR} contenant $\overline{K}$ et B_{st}. Ce théorème revient à dire que, pour toute représentation de de Rham V, l'inclusion $(\overline{K}B_{st}) \otimes_{\mathbb{Q}_p} V)^{G_K} \subset (B_{dR} \otimes_{\mathbb{Q}_p} V)^{G_K}$ est une égalité. Berger [Be02] en a ramené la preuve à un résultat sur les équations différentielles p-adiques, résultat qui a ensuite été prouvé indépendamment par André [An02], Kedlaya [Ke02] et Mebkhout [Me02], voir §3.

L'un des intérêts de ce théorème est que l'on dispose d'une description algébrique *explicite* de la catégorie des représentations potentiellement semi-stables. Le Frobenius usuel sur $W(R)$ s'étend de façon naturelle en un endomorphisme φ de l'anneau B_{st} (on a $\varphi t = pt$ et $\varphi(\log[\pi]) = p\log[\pi]$). Il existe une unique B_{cris}-dérivation $N : B_{st} \to B_{st}$ telle que $N(\log[\pi]) = -1$. L'action de φ et de N commutent à celle de G_K et $N\varphi = p\varphi N$.

Soit $L \in \mathcal{L}$ telle que L/K est galoisienne. Pour toute représentation p-adique V de G_K, $D_{st,L}(V)$ est un $(\varphi, N, \mathrm{Gal}(L/K))$-*module filtré* de dimension finie, i.e. c'est un L_0-espace vectoriel D de dimension finie, muni

– de deux applications $\varphi : D \to D$, $N : D \to D$, la première semi-linéaire relativement à la restriction de φ à L_0 et bijective, la deuxième linéaire, vérifiant $N\varphi = p\varphi N$,

– d'une action semi-linéaire de $\mathrm{Gal}(L/K)$, commutant à φ et à N,

– d'une filtration indexée par $\mathbb{Z}$, décroissante, exhaustive et séparée, du K-espace vectoriel $D_K = (L \otimes_{L_0} D)^{\mathrm{Gal}(L/K)}$ (si $D = D_{st,L}(V)$, on a $D_K \subset (B_{dR} \otimes_{\mathbb{Q}_p} V)^{G_K}$ et, pour tout $i \in \mathbb{Z}$, $\mathrm{Fil}^i D_K = D_K \cap (B_{dR}^+ t^i \otimes_{\mathbb{Q}_p} V)^{G_K}$).

On pose $t_H(D) = \sum_{i\in\mathbb{Z}} i.\dim_K \mathrm{Fil}^i D_K/\mathrm{Fil}^{i+1} D_K$. Si $D = \oplus_{\alpha\in\mathbb{Q}} D_\alpha$ est la décomposition suivant les pentes du φ-isocristal sous-jacent, on pose aussi $t_N(D) = \sum_{\alpha\in\mathbb{Q}} \alpha.\dim_{L_0} D_\alpha$. On dit que D est *admissible* si

a) on a $t_H(D) = t_N(D)$,

b) pour tout sous-L_0-espace vectoriel D' de D, stable par φ, N et $\mathrm{Gal}(L/K)$, on a $t_H(D') \leq t_N(D')$ (on a muni $D'_K \subset D_K$ de la filtration induite).

THÉORÈME C [CF00]. — *Soit $L \subset \overline{K}$ une extension finie galoisienne de K.*

i) Pour toute représentation L-semi-stable V, $D_{st,L}(V)$ est admissible.

ii) Le foncteur qui à V associe $D_{st,L}(V)$ induit une équivalence(2) *entre la sous-catégorie pleine* $\mathrm{Rep}_{st,L}(G_K)$ *de* $\mathrm{Rep}_{\mathbb{Q}_p}(G_K)$ *dont les objets sont les représentations L-semi-stables et la catégorie des $(\varphi, N, \mathrm{Gal}(L/K))$-modules filtrés admissibles.*

Remarque. Il était jusqu'à présent d'usage [Fo88b] d'appeler *faiblement admissible* ce que nous appelons ici *admissible*. On savait (*loc.cit.*, th.5.6.7) que $D_{st,L}$ induit une équivalence entre la catégorie $\mathrm{Rep}_{st,L}(G_K)$ et une sous-catégorie pleine de la catégorie des modules filtrés (faiblement) admissibles ; on conjecturait que ce foncteur est essentiellement surjectif et c'est ce qui est prouvé dans [CF00].

2. Espaces de Banach-Colmez(3)

2.1. – Une *C-algèbre de Banach* est une C-algèbre normée complète A ; son *spectre maximal* est l'ensemble $\mathrm{Spm}_C A$ des sections continues $s : A \to C$ du morphisme structural. Si $f \in A$ et $s \in \mathrm{Spm}_C A$, on pose $f(s) = s(f)$. Une *C-algèbre spectrale* est une C-algèbre de Banach A telle que la norme est la norme spectrale, i.e. telle que, pour tout $f \in A$, $||f|| = sup_{s \in \mathrm{Spm}_C A} |f(s)|_p$; dans ce cas, $\mathrm{Spm}_C A$ est un espace métrique complet (la distance étant définie par $d(s_1, s_2) = \sup_{||f|| \leq 1} |f(s_1) - f(s_2)|_p$). Avec comme morphismes les homomorphismes continus de C-algèbres, les C-algèbres spectrales forment une catégorie. La *catégorie des variétés spectrales affines sur C* est la catégorie opposée.

Un *groupe spectral commutatif affine sur C* est un objet en groupes commutatifs dans la catégorie des variétés spectrales affines sur C. Ces groupes forment, de façon évidente, une catégorie additive qui a des limites projectives finies. Le foncteur qui à un groupe spectral commutatif affine associe le groupe topologique sous-jacent est fidèle. Si $\mathcal{S} = \mathrm{Spm}_C A$ est un groupe spectral commutatif affine, *un sous-groupe spectral* du groupe topologique sous-jacent est un sous-groupe $\mathcal{T}$ qui admet une structure de groupe spectral (nécessairement unique) telle que l'inclusion $\mathcal{T} \to \mathcal{S}$ est un morphisme de groupes spectraux.

Soit S un espace de Banach (p-adique) et $\mathcal{S}_0$ la boule unité. Un *réseau* de S est un sous-$\mathbb{Z}_p$-module $\mathcal{S}$ qui est tel que l'on peut trouver $r, s \in \mathbb{Z}$ vérifiant

(2) C'est même une équivalence de catégories tannakiennes, cf. [Fo88b].

(3) C'est en cherchant à prouver le théorème C que j'ai été conduit à m'intéresser aux presque C-représentations. C'est Colmez qui a compris que les propriétés dont j'avais besoin provenaient de structures analytiques. Cela nous a permis de prouver le théorème C. Colmez a ensuite étudié plus en détail ces structures analytiques [Co02]. Ce que je raconte ici est une interprétation, dans le langage de [Fo02], §4, de ces travaux de Colmez et devrait être développé dans [FP02].

$p^r\mathcal{S}_0 \subset \mathcal{S} \subset p^s\mathcal{S}_0$. Il revient au même de dire qu'il existe une norme équivalente à la norme donnée pour laquelle $\mathcal{S}$ est la boule unité.

Une *C-structure analytique sur S* est la donnée d'un C-groupe spectral commutatif affine $\mathcal{S}$ et d'un homomorphisme continu du groupe topologique sous-jacent à $\mathcal{S}$ dans S dont l'image est un réseau et le noyau un $\mathbb{Z}_p$-module de type fini. On dit que deux C-structures analytiques $\mathcal{S}$ et $\mathcal{T}$ sur S sont *équivalentes* si $\mathcal{S} \times_S \mathcal{T}$ est un sous-groupe spectral de $\mathcal{S} \times \mathcal{T}$. Un *(espace de) Banach analytique (sur C)* est la donnée d'un espace de Banach muni d'une classe d'équivalence de C-structures analytiques (on les appelle les structures *admissibles* de S). On dit que S est *effectif* s'il existe une structure admissible $\mathcal{S}$ telle que l'application $\mathcal{S} \to S$ est injective.

Un *morphisme de Banach analytiques* $f : S \to T$ est une application $\mathbb{Q}_p$-linéaire continue telle qu'il existe des structures admissibles $\mathcal{S}$ de S et $\mathcal{T}$ de T et un morphisme $\mathcal{S} \to \mathcal{T}$ qui induit f. Les Banach analytiques forment une catégorie additive $\mathcal{BA}_C$.

Si S est un Banach analytique et si V est un sous-$\mathbb{Q}_p$-espace vectoriel de dimension finie, le quotient S/V a une structure naturelle de Banach analytique. On dit que deux Banach analytiques S_1 et S_2 sont *presque isomorphes* s'il existe des sous-$\mathbb{Q}_p$-espaces vectoriels de dimension finie V_1 de S_1 et V_2 de S_2 et un isomorphisme $S_1/V_1 \to S_2/V_2$ (de Banach analytiques).

Le groupe sous-jacent à $\mathcal{O}_C$ a une structure naturelle de groupe spectral commutatif affine : on a $\mathcal{O}_C = \mathrm{Spm}_C C\{X\}$ où $C\{X\}$ est l'algèbre de Tate des séries formelles à coefficients dans C en l'indéterminée X dont le terme général tend vers 0. Ceci fait de C un espace de Banach analytique effectif. Un *Banach analytique vectoriel* est un Banach analytique isomorphe à C^d pour un entier d convenable. Un *espace de Banach-Colmez* est un *Banach analytique presque vectoriel*, i.e. un Banach analytique qui est presque isomorphe à un Banach analytique vectoriel. On note $\mathcal{BC}_C$ la sous-catégorie pleine de $\mathcal{BA}_C$ dont les objets sont les espaces de Banach-Colmez.

Proposition (théorème de Colmez [4]). — *La catégorie $\mathcal{BC}_C$ est abélienne et le foncteur d'oubli de $\mathcal{BC}_C$ dans la catégorie des $\mathbb{Q}_p$-espaces vectoriels est exact et fidèle. Il existe sur les objets de $\mathcal{BC}_C$, une unique fonction additive $dh : S \mapsto (d(S), h(S)) \in \mathbb{N} \times \mathbb{Z}$ telle que $dh(C^d) = (d, 0)$ et $dh(V) = (0, \dim_{\mathbb{Q}_p} V)$ si V est de dimension finie sur $\mathbb{Q}_p$.*

2.2. – La meilleure façon de comprendre les théorèmes A et C c'est d'utiliser le résultat précédent pour les prouver [5]. Lorsque k est fini, toute presque-C-représen-

[4] C'est à peu près le résultat principal de [Co02]. La définition donnée par Colmez de ce qu'il appelle les Espaces de Banach de dimension finie (avec un E majuscule) est légèrement différente. Il n'est pas très difficile de construire une équivalence entre sa catégorie et la nôtre [FP02].

[5] C'est ainsi que Colmez redémontre le théorème C dans [Co02]. Moyennant une preuve un peu plus compliquée, on peut n'utiliser qu'un résultat d'analyticité apparemment moins fort ; c'est ce qu'on fait pour prouver le théorème C dans [CF00] et le théorème A dans [Fo02].

tation est munie d'unee structure naturelle d'espace de Banach-Colmez ; toute application $\mathbb{Q}_p$-linéaire continue G_K-équivariante d'une presque C-représentation dans une autre est analytique. Le fait que $\mathcal{C}(G_K)$ est abélienne et l'existence de la fonction dh résultent alors du théorème de Colmez.

Le principe de la preuve du théorème C est le suivant : On se ramène facilement au cas semi-stable, i.e. au cas où $L = K$. Il s'agit de vérifier que, si D est un (φ, N)-module filtré (faiblement) admissible de dimension h, il existe une représentation p-adique V de dimension h telle que $D_{st,K}(V)$ soit isomorphe à D. Une torsion à la Tate permet de supposer que $\mathrm{Fil}^0 D_K = D_K$. Notons $V_{st}^{+,0}(D)$ le $\mathbb{Q}_p$-espace vectoriel des applications K_0-linéaires de D dans $B_{st} \cap B_{dR}^+$ qui commutent à l'action de φ et de N et $V_{st}^{+,1}$ le quotient du K-espace vectoriel des applications K-linéaires de D_K dans B_{dR}^+ par le sous-espace des applications qui sont compatibles avec la filtration. On commence par vérifier que le noyau $V_{st}^*(D)$ de l'application évidente $\beta : V_{st}^{+,0}(D) \to V_{st}^{+,1}(D)$ est un $\mathbb{Q}_p$-espace vectoriel de dimension finie $\leq h$ et que, s'il est de dimension h, alors la représentation duale $V_{st}(D)$ est semi-stable et D est isomorphe à $D_{st}(V_{st}(D))$. La théorie des espaces de Banach analytiques permet de munir $V_{st}^{+,0}(D)$ et $V_{st}^{+,1}(D)$ d'une structure d'espace de Banach-Colmez et on a $dh(V_{st}^{+,0}(D)) = (t_N(D), h)$ tandis que $dh(V_{st}^{+,1}(D)) = (t_H(D), 0)$. Il suffit alors de vérifier que l'application β est analytique. Comme $t_H(D) = t_N(D)$, l'additivité de dh implique que β est surjective et que $dh(V_{st}^*(D)) = (0, h)$, ce qui signifie bien que $\dim_{\mathbb{Q}_p} V_{st}^*(D) = h$.

3. Equations différentielles

3.1. – Soit A un anneau commutatif et $d : A \to \Omega_A$ une dérivation de A dans un A-module Ω_A. Ici, un A-module à connexion (sous-entendu relativement à d) est un A-module libre de rang fini $\mathcal{D}$ muni d'une application $\nabla : \mathcal{D} \to \mathcal{D} \otimes \Omega_A$ vérifiant la règle de Leibniz. On dit que ce module est *trivial* s'il est engendré par le sous-groupe $\mathcal{D}_{\nabla=0}$ des *sections horizontales.*

Pour tout corps L de caractéristique 0, complet pour une valuation discrète, notons (cf. par exemple [Ts98], §2) $\mathcal{R}_{x,L}$ l'*anneau de Robba de* L (ou "anneau des fonctions analytiques sur une couronne d'épaisseur nulle"), c'est-à-dire l'anneau des séries $\sum_{n\in\mathbb{Z}} a_n x^n$ à coefficients dans L vérifiant

$\forall s < 1, |a_n| s^n \mapsto 0$ si $n \mapsto +\infty$ et $\exists r < 1$ tel que $|a_n| r^n \mapsto 0$ si $n \mapsto -\infty$.

Le sous-anneau $\mathcal{E}_{x,L}^\dagger$ de $\mathcal{R}_{x,L}$ des fonctions $\sum a_n x^n$ telles que les a_n sont bornés est un corps muni d'une valuation discrète (définie par $|\sum a_n x^n| = \sup |a_n|$) qui n'est pas complet mais est hensélien. Son corps résiduel s'identifie au corps des séries formelles $E = k_L((x))$ où k_L désigne le corps résiduel de L. Pour toute extension finie séparable F de E, il existe une, unique à isomorphisme unique près, extension non ramifiée $\mathcal{E}_F^\dagger$ de $\mathcal{E}_{x,L}^\dagger$ de corps résiduel F . Posons $\mathcal{R}_F = \mathcal{R}_{x,L} \otimes_{\mathcal{E}_{x,L}^\dagger} \mathcal{E}_F^\dagger$. Si k_F désigne le corps résiduel de F, L' l'unique extension non ramifiée de L de corps résiduel k_F et si x' est un relèvement dans l'anneau des entiers de $\mathcal{E}_F^\dagger$ d'une

uniformisante de F, l'anneau $\mathcal{R}_F$ s'identifie à l'anneau de Robba $\mathcal{R}_{x',L'}$.

Notons $\Omega^1_{\mathcal{R}_{x,L}}$ le $\mathcal{R}_{x,L}$-module libre de rang 1 de base dx, solution du problème universel pour les dérivations continues en un sens évident. Les modules à connexion sur l'anneau $\mathcal{R}_{x,L}$ forment une catégorie artinienne. Si $\mathcal{D}$ est un objet de cette catégorie, on dit qu'il est *unipotent* si son semi-simplifié est trivial. On dit qu'il est *quasi-unipotent* s'il existe une extension finie séparable F de $k((X))$ telle que le module à connexion sur $\mathcal{R}_F$ déduit de $\mathcal{D}$ par extension des scalaires soit unipotent.

Pour tout z dans l'aneeau des entiers de $\mathcal{E}^{\dagger}_{x,K_0}$, il existe un unique endomorphisme continu φ de $\mathcal{R}_{x,K_0}$ qui prolonge le Frobenius absolu sur K_0 et vérifie $\varphi(x) = x^p + pz$; on appelle *Frobenius* un tel endomorphisme. Pour un tel φ, on note encore $\varphi : \Omega^1_{\mathcal{R}_{x,K_0}} \to \Omega^1_{\mathcal{R}_{x,K_0}}$ l'application induite. Soit $\mathcal{D}$ un module à connexion sur $\mathcal{R}_{x,K_0}$. Une *structure de Frobenius sur* $\mathcal{D}$ consiste en la donnée d'un Frobenius φ sur $\mathcal{R}_{x,K_0}$ et d'une application φ-semi-linéaire $\varphi_{\mathcal{D}} : \mathcal{D} \to \mathcal{D}$ commutant à ∇.

THÉORÈME (André, Kedlaya, Mebkhout [6]). — *Tout module à connexion sur $\mathcal{R}_{x,K_0}$ qui admet une structure de Frobenius est quasi-unipotent.*

Avant de montrer comment Berger [Be2] déduit le théorème B de cet énoncé, rappelons quelques résultats de [Fo00], [Fo90] et [CC98] (cf. aussi [Co98]). Dans tout ce qui suit, V est une représentation p-adique de G_K de dimension finie h.

3.2. – Soit K_∞ le sous-corps de $\overline{K}$ engendré sur K par les racines de l'unité d'ordre une puissance de p. Posons $H_K = \mathrm{Gal}(\overline{K}/K_\infty)$ et $\Gamma_K = G_K/H_K$. En utilisant la théorie de Sen [Se80], on montre [Fo00] que l'union $\Delta_{dR}(V)$ des sous-$K_\infty[[t]]$-modules de type fini de $(B_{dR} \otimes_{\mathbb{Q}_p} V)^{H_K}$ stables par Γ_K est un $K_\infty((t))$-espace vectoriel de dimension h et qu'il existe une unique connexion

$$\nabla : \Delta_{dR}(V) \to \Delta_{dR}(V) \otimes dt/t$$

qui a la propriété que, pour tout sous-$K_\infty[[t]]$-module de type fini Y stable par G_K, tout entier $r \geq 0$ et tout $y \in Y$, il existe un sous-groupe ouvert $\Gamma_{r,y}$ de Γ tel que, si $\nabla(y) = \nabla_0(y) \otimes dt/t$, alors

$$\gamma(y) \equiv \exp(\log \chi(\gamma).\nabla_0)(y) \pmod{t^r Y}, \text{ pour tout } \gamma \in \Gamma_{r,y}.$$

Cette connexion est *régulière* : le $K_\infty[[t]]$-module $\Delta^+_{dR}(V) = (B^+_{dR} \otimes V) \cap \Delta_{dR}(V)$ est un réseau de $\Delta_{dR}(V)$ vérifiant $\nabla(\Delta^+_{dR}(V)) \subset \Delta^+_{dR}(V) \otimes dt/t$. On a $D_{dR}(V) = (\Delta_{dR}(V))^{\Gamma_K}$. L'action de Γ_K est discrète sur $\Delta_{dR}(V)_{\nabla=0}$; on en déduit que $\Delta_{dR}(V)_{\nabla=0} = K_\infty \otimes_K D_{dR}(V)$ donc que V est de de Rham si et seulement si le module à connexion $\Delta_{dR}(V)$ est trivial. Ceci se produit si et seulement s'il existe un réseau (nécessairement unique) $\Delta^0_{dR}(V)$ de $\Delta_{dR}(V)$ vérifiant $\nabla(\Delta^0_{dR}(V)) \subset \Delta^0_{dR}(V) \otimes dt$.

[6] Crew [Cr98] a suggéré que ce théorème pouvait être vrai ; il a été prouvé indépendamment par André [An02], Mebkhout [Me02] et Kedlaya [Ke01]. Pour André comme pour Mebkhout, c'est un cas particulier d'un résultat plus général dont la preuve repose sur la théorie de Christol-Mebkhout [CM]. La preuve de Kedlaya est plus directe : elle utilise une classification à la Dieudonné-Manin des modules munis d'un Frobenius pour se ramener à un résultat de Tsuzuki [Ts98]. Voir [Co01] pour une étude comparative plus détaillée.

3.3. – Rappelons brièvement la théorie des (φ,Γ)-modules [Fo80]. Soit $\mathcal{O}_{\mathcal{E}_0}$ l'adhérence dans $W(FR)$ de la sous-$W(k)$-algèbre engendrée par $[\varepsilon]$ et $1/([\varepsilon]-1)$. C'est un anneau de valuation discrète complet dont l'idéal maximal est engendré par p et dont le corps résiduel E_0 est le corps des séries formelles $k((\varepsilon-1))$ vu comme sous-corps fermé de FR. Notons $\mathcal{O}_{\widehat{\mathcal{E}}^{nr}}$ le séparé complété pour la topologie p-adique de l'union de toutes les sous-$\mathcal{O}_{\mathcal{E}_0}$-algèbres finies étales de $\mathcal{O}_{\mathcal{E}_0}$ contenues dans $W(FR)$. C'est un anneau de valuation discrète complet dont le corps résiduel est une clôture séparable E^s de E_0. Son corps des fractions $\widehat{\mathcal{E}}^{nr}$ s'identifie à un sous-corps fermé du corps $\widetilde{B}=W(FR)[1/p]$, stable par l'action de G_K et par le Frobenius φ. Le corps $\mathcal{E}_K=(\widehat{\mathcal{E}}^{nr})^{H_K}$ est une extension finie non ramifiée du corps des fractions de $\mathcal{O}_{\mathcal{E}_0}$. Son corps résiduel E_K est une extension finie séparable de E_0 ; le corps résiduel k' de E_K est celui de K_∞.

Alors, $D(V)=(\widehat{\mathcal{E}}^{nr}\otimes_{\mathbb{Q}_p}V)^{H_K}$ est un (φ,Γ_K)-*module sur* $\mathcal{E}_K$, i.e. un $\mathcal{E}_K$-espace vectoriel de dimension finie D muni d'un Frobenius φ-semi-linéaire (que l'on note encore φ) et d'une action semi-linéaire continue de Γ_K commutant à l'action de φ ; ce (φ,Γ_K)-module est *étale*, i.e. il existe un $\mathcal{O}_{\mathcal{E}_K}$-réseau $\mathcal{D}$ de D tel que $\mathcal{D}$ est le $\mathcal{O}_{\mathcal{E}_K}$-module engendré par $\varphi(\mathcal{D})$. La correspondance $V\mapsto D(V)$ définit une équivalence entre $\mathrm{Rep}_{\mathbb{Q}_p}(G_K)$ et la catégorie des (φ,Γ_K)-modules étales.

3.4. – Il n'y a pas de flèche naturelle de $\widehat{\mathcal{E}}^{nr}$ dans B_{dR}^+, ce qui fait que la comparaison entre $\Delta_{dR}(V)$ et $D(V)$ n'est pas si facile. Toutefois, si $a\in R$ est non nul, $[a]\in W(R)\subset B_{dR}^+$ est inversible dans B_{dR}^+, ce qui permet de voir $1/[a]=[1/a]$ comme un élément de B_{dR}^+. Tout élément de $\widetilde{B}$ s'écrit d'une manière et d'une seule sous la forme $\sum_{n\gg-\infty}p^n[a_n]$, avec les $a_n\in FR$; notons $\widetilde{B}_{dR}^+$ le sous-anneau de $\widetilde{B}$ formé des séries de ce type qui convergent dans B_{dR}^+. L'application $\widetilde{B}_{dR}^+\to B_{dR}^+$ est injective et permet d'identifier $\widetilde{B}_{dR}^+$ à une sous-$W(R)[1/p]$-algèbre de B_{dR}^+. Pour tout $r\in\mathbb{N}$, posons $\widehat{\mathcal{E}}_r^{nr,\dagger}=\widehat{\mathcal{E}}^{nr}\cap\varphi^r(\widetilde{B}_{dR}^+)$ et, pour tout $b\in\widehat{\mathcal{E}}_r^{nr,\dagger}$, notons $\varphi_r(b)$ l'unique $c\in\widetilde{B}_{dR}^+\subset B_{dR}^+$ tel que $\varphi^r(c)=b$. On a $\widehat{\mathcal{E}}_r^{nr,\dagger}\subset\widehat{\mathcal{E}}_{r+1}^{nr,\dagger}$; soit $\widehat{\mathcal{E}}^{nr,\dagger}$ l'union des $\widehat{\mathcal{E}}_r^{nr,\dagger}$. Alors $\mathcal{E}_K^\dagger=(\widehat{\mathcal{E}}^{nr,\dagger})^{H_K}$ est un sous-corps dense de $\mathcal{E}_K$ stable par φ. On pose $D^\dagger(V)=(\widehat{\mathcal{E}}^{nr,\dagger}\otimes_{\mathbb{Q}_p}V)^{H_K}$. On peut le calculer à partir de $D(V)$: c'est l'union des sous-$\mathcal{E}_K^\dagger$-espaces vectoriels de dimension finie de $D(V)$ stables par φ. Le résultat principal de [CC98] est que V est surconvergente, c'est-à-dire que l'application naturelle $\mathcal{E}_K\otimes D^\dagger(V)\to D(V)$ est un isomorphisme.

Pour tout $r\in\mathbb{N}$, soit $\mathcal{E}_{K,r}^\dagger=(\widehat{\mathcal{E}}_r^{nr,\dagger})^{H_K}$. Alors $D_r^\dagger(V)=(\widehat{\mathcal{E}}_r^{nr,\dagger}\otimes_{\mathbb{Q}_p}V)^{H_K}$ est aussi le plus grand sous-$\mathcal{E}_{K,r}^\dagger$-module M de type fini de $D(V)$ tel que $\varphi(M)\subset\mathcal{E}_{K,r+1}^\dagger M$. Pour r assez grand, l'application naturelle $\mathcal{E}_K^\dagger\otimes_{\mathcal{E}_{K,r}^\dagger}D_r^\dagger(V)\to D^\dagger(V)$ est un isomorphisme. Lorsqu'il en est ainsi, on a $\varphi_r(\mathcal{E}_{K,r}^\dagger)\subset K_\infty[[t]]$ et

$$\Delta_{dR}(V)=K_\infty((t))\otimes_{\mathcal{E}_{K,r}^\dagger}D_r^\dagger(V)\text{ et donc }D_{dR}(V)=(K_\infty((t))\otimes_{\mathcal{E}_{K,r}^\dagger}D_r^\dagger(V))^{\Gamma_K}.$$

3.5. – Wach [Wa96] a montré comment calculer $D_{st}(V)$ à partir de $D^\dagger(V)$ lorsque

V est *de hauteur finie.* C'est Berger [Be02] qui a compris comment traiter le cas général : Choisissons un relèvement x dans l'anneau des entiers de $\mathcal{E}_K^\dagger$ d'une uniformisante de E_K. Si K_0' désigne le corps des fractions de $W(k')$, $\mathcal{E}_K^\dagger$ s'identifie précisément au sous-anneau $\mathcal{E}_{x,K_0'}^\dagger$ de l'anneau de Robba $\mathcal{R}_{x,K_0'}$. Ce dernier ne dépend pas du choix de x et nous le notons $\mathcal{E}_K^{rig}$; il contient $t = \log[\varepsilon]$. Si L est une extension finie de K contenue dans $\overline{K}$, le corps $F = (E^s)^{\mathrm{Gal}(\overline{K}/LK_\infty)}$ est une extension finie de E_K et l'anneau $\mathcal{E}_L^{rig}$ s'identifie à l'anneau noté $\mathcal{R}_F$ au §3.1.

Rappelons (§1.2) que si $u = \log[\varepsilon - 1]$, on a $B_{st} = B_{cris}[u]$. Les actions de φ et de Γ_K s'étendent de façon évidente à $\mathcal{E}_K^{rig}$, à $\mathcal{E}_K^{rig}[1/t]$ et à l'anneau $\mathcal{E}_K^{rig}[1/t][u]$ des polynômes en u à coefficients dans $\mathcal{E}_K^{rig}[1/t]$. Berger montre que

$$D_{cris}(V) = (\mathcal{E}_K^{rig}[1/t] \otimes_{\mathcal{E}_K^\dagger} D^\dagger(V))^{\Gamma_K} \text{ et } D_{st}(V) = (\mathcal{E}_K^{rig}[1/t][u] \otimes_{\mathcal{E}_K^\dagger} D^\dagger(V))^{\Gamma_K}$$

(l'action de N sur $D_{st}(V)$ est la restriction de $-d/du \otimes id_{D^\dagger(V)}$).

3.6. – Posons $D = D_K^{rig}(V) = \mathcal{E}_K^{rig}[1/t] \otimes_{\mathcal{E}_K^\dagger} D^\dagger(V)$. En utilisant l'action de Γ_K comme au §3.4, on définit une connexion $\nabla : D \to D \otimes dt/t$ qui commute à l'action de φ. Cette connexion est *régulière* au sens qu'il existe un sous-$\mathcal{E}_K^{rig}$-module D^+ de D, libre de rang h, stable par φ et vérifiant $\nabla(D^+) \subset D^+ \otimes dt/t$ (prendre $D^+ = \mathcal{E}_K^{rig} \otimes_{\mathcal{E}_K^\dagger} D^\dagger(V)$). On vérifie que le $\mathcal{E}_K^{rig}$-module libre $\Omega^1_{\mathcal{E}_K^{rig}}$ admet $d[\varepsilon]$ comme base. Mais $dt/t = [\varepsilon]^{-1}/td[\varepsilon]$ et t n'est pas inversible dans $\mathcal{E}_K^{rig}$. On déduit alors facilement du théorème d'André-Kedlaya-Mebkhout que V est potentiellement semi-stable si et seulement s'il existe un sous-$\mathcal{E}_K^{rig}$-module libre D^0 de D, libre de rang h, stable par φ et vérifiant $\nabla(D^0) \subset D^0 \otimes dt$.

Il ne reste plus qu'à construire un tel D^0 lorsque V est de de Rham. Fixons un entier $r_0 \geq 1$ suffisamment grand pour que $D_{r_0}^\dagger(V)$ contienne une base de $D^\dagger(V)$ sur $\mathcal{E}_K^\dagger$ et pour que $x \in \mathcal{E}_{r_0}^\dagger$. Pour tout $r \geq r_0$ le sous-anneau $\mathcal{E}_{K,r}^{rig}$ de $\mathcal{E}_K^{rig} = \mathcal{R}_{x,K_0'}$ formé des $\sum_{n\in\mathbb{Z}} a_n x^n$ vérifiant

$$\forall s < 1, |a_n| s^n \mapsto 0 \text{ si } n \mapsto +\infty \text{ et } a_n(\varepsilon^r - 1)^n \mapsto 0 \text{ si } n \mapsto -\infty$$

est stable par Γ_K et contient $\mathcal{E}_{K,r}^\dagger$. Si $D_r = D_{K,r}^{rig}(V) = \mathcal{E}_{K,r}^{rig} \otimes_{\mathcal{E}_{K,r}^\dagger} D_r^\dagger(V)$, alors D est la réunion croissante des D_r et $\varphi(D_r) \subset D_{r+1}$. L'application φ_r induit un homomorphisme de $\mathcal{E}_{K,r}^{rig}$ dans $K_\infty((t))$ et un isomorphisme de $K_\infty((t)) \otimes_{\mathcal{E}_{K,r}^{rig}} D_r$ sur $\Delta_{dR}(V)$. L'application $\Phi_r : D_r \to \Delta_{dR}(V)$ qui envoie a sur $1 \otimes a$ est injective.

Soit $\Delta_{dR}^0(V)$ le sous-$K_\infty((t))$-module de $\Delta_{dR}(V)$ engendré par les sections horizontales. Pour tout $r \geq r_0$, soit $D_r^0 = \{a \in D_r \mid \Phi_s(a) \in \Delta_{dR}^0(V) \text{ pour tout } s \geq r\}$. On a $D_r^0 \subset D_{r+1}^0$ et $D^0 = \cup_{r\geq r_0} D_r^0$ est un sous-$\mathcal{E}_K^{rig}$-module de D, stable par φ et vérifiant $\nabla(D^0) \subset D^0 \otimes dt$. Si V est de de Rham on déduit du fait que $\Delta_{dR}^0(V)$ est un réseau de $\Delta_{dR}(V)$ que D^0 est libre de rang h sur $\mathcal{E}_K^{rig}$. D'où le théorème B.

Bibliographie

[An02] Y. André, *Filtration de type Hasse-Arf et monodromie p-adique*, Inv. Math. **148** (2002), 285–317.

[Be02] L. Berger, *Représentations p-adiques et équations différentielles*, Inv. Math. **148** (2002), 219-284.

[CM] G. Christol et Z. Mebkhout, *Sur le théorème de l'indice des équations différentielles p-adiques* I, Ann. Inst. Fourier **43** (1993), 1545–1574 ; II, Ann. of Maths. **146** (1997), 345–410 ; III, Ann. of Maths. **151** (2000), 385–457 ; IV, Inv. Math. **143** (2001), 629–671.

[Co98] P. Colmez *Représentations p-adiques d'un corps local* in Proceedings of the I.C.M. Berlin, vol. II, Documenta Mathematica (1998), 153–162.

[Co01] P. Colmez, *Les conjectures de monodromie p-adique*, Sém. Bourbaki, exp. 897, novembre 2001.

[Co02] P. Colmez, *Espaces de Banach de dimension finie*, J. Inst. Math. Jussieu, à paraître.

[CC98] F. Cherbonnier et P. Colmez, *Représentations p-adiques surconvergentes*, Inv. Math. **133** (1998), 581–611.

[CF00] P. Colmez et J.-M. Fontaine, *Construction des représentations semi-stables*, Inv. Math. **140** (2000), 1–43.

[Cr98] R. Crew, *Finiteness theorems for the cohomology of an overconvergent isocrystal on a curve*, Ann. scient. E.N.S. **31** (1998), 717–763.

[Fo83] J.-M. Fontaine, *Représentations p-adiques*, in Proceedings of the I.C.M., Warszawa, vol. I, Elsevier, Amsterdam (1984), 475–486.

[Fo88a] J.-M. Fontaine, *Le corps des périodes p-adiques*, avec un appendice par Pierre Colmez, in Périodes p-adiques, Astérisque **223**, S.M.F., Paris (1994), 59–111.

[Fo88b] J.-M. Fontaine, *Représentations p-adiques semi-stables*, in Périodes p-adiques, Astérisque **223**, S.M.F., Paris (1994), 113–184.

[Fo90] J.-M. Fontaine, *Représentations p-adiques des corps locaux*, in the Grothendieck Festschrift, vol II, Birkhäuser, Boston (1990), 249–309).

[Fo00] J.-M. Fontaine, *Arithmétique des représentations galoisiennes p-adiques*, prépublication, Orsay 2000-24. A paraître dans Astérisque.

[Fo02] J.-M. Fontaine, *Presque-$\mathbb{C}_p$-représentations*, prépublication, Orsay 2002-12.

[FP02] J.-M. Fontaine et Jérôme Plût, *Espaces de Banach-Colmez*, en préparation.

[Ke02] K. Kedlaya, *A p-adic local monodromy theorem*, preprint, Berkeley (2001).

[Me02] Z. Mebkhout, *Analogue p-adique du théorème de Turrittin et le théorème de la monodromie p-adique*, Inv. Math. **148** (2002), 319–351.

[Sen80] S.Sen, Continuous Cohomology and p-adic Galois Representations, Inv. Math. **62** (1980), 89–116.

[Ts98] N. Tsuzuki, *Slope filtration of quasi-unipotent overconvergent F-isocrystals*, Ann. Inst. Fourier **48** (1998), 379–412.

[Wa96] N. Wach, *Représentations p-adiques potentiellement cristallines*, Bull. S.M.F. **124** (1996), 375–400.

ICM 2002 · Vol. II · 149–162

Equivariant Bloch-Kato Conjecture and Non-abelian Iwasawa Main Conjecture

A. Huber* G. Kings†

Abstract

In this talk we explain the relation between the (equivariant) Bloch-Kato conjecture for special values of L-functions and the Main Conjecture of (non-abelian) Iwasawa theory. On the way we will discuss briefly the case of Dirichlet characters in the abelian case. We will also discuss how "twisting" in the non-abelian case would allow to reduce the general conjecture to the case of number fields. This is one the main motivations for a non-abelian Main Conjecture.

2000 Mathematics Subject Classification: 11G40, 11R23, 19B28.
Keywords and Phrases: Iwasawa theory, L-function, Motive.

1. Introduction

The class number formula expresses the leading coefficient of a Dedekind-ζ-function of a number field F in terms of arithmetic invariants of F:

$$\zeta_F(0)^* = -\frac{hR_F}{w_F}$$

(h the class number, R_F the regulator, w_F the number of roots of unity in F). By work of Lichtenbaum, Bloch, Beilinson, and Kato among others, the class number formula has been generalized to other L-functions of varieties (or even motives) culminating in the Tamagawa number conjecture by Bloch and Kato.

Iwasawa, on the other hand, initiated the study of the growth of the class numbers in towers of number fields. His decisive idea was to consider the class group of the tower as a module under the completed group ring of the Galois group of the tower. From his work evolved the "Main Conjecture" describing this growth in terms of the p-adic L-function.

*Math. Institut, Universität Leipzig, Augustusplatz 10/11, 04109 Leipzig, Germany. E-mail: huber@mathematik.uni-leipzig.de

†NWF I-Mathematik, Universität Regensburg, 93040 Regensburg, Germany. E-mail: guido.kings@mathematik.uni-regensburg.de

It is a surprising insight of Kato that an equivariant version of the Tamagawa number conjecture can be viewed as a version of the Main Conjecture of Iwasawa theory. Perrin-Riou, in her efforts to develop a theory of p-adic L-functions, arrived at a similar conclusion.

The purpose of this paper is to make the connection between the equivariant Tamagawa number conjecture and the Iwasawa Main Conjecture precise. In the spirit of Kato, we formulate an Iwasawa Main Conjecture (3.2.1) for arbitrary motives and towers of number fields whose Galois group is a p-adic Lie group. This formulation does not involve p-adic L-functions. We show that it is implied by the equivariant Tamagawa number conjecture formulated by Burns and Flach. For ease of exposition, we restrict to the case of L-values at very negative integers, where the Bloch-Kato exponential does not play a role. The study of non-abelian Iwasawa theory was initiated by Coates. Recently, there have been systematic results by Coates, Howson, Ochi, Schneider, Sujatha and Venjakob.

Our interest in allowing general towers of number fields is motivated by the possibility of reducing the Tamagawa number conjecture to an equivariant class number formula (modulo hard conjectures, see 3.).

Important special cases of the Main Conjecture were considered by (alphabetical order) Coates, Greenberg, Iwasawa, Kato, Mazur, Perrin-Riou, Rubin, Schneider, Wiles and more recently by Ritter and Weiss.

It is a pleasure to thank C. Deninger, S. Howson, B. Perrin-Riou, A. Schmidt, P. Schneider for helpful comments and discussions.

2. Non-abelian equivariant Tamagawa number conjecture

2.1. Notation

Fix $p \neq 2$ and let M be a motive over $\mathbb{Q}$ with coefficients in $\mathbb{Q}$, for example $M = h^r(X)$, X a smooth projective variety over $\mathbb{Q}$. It has Betti-realization M_B and p-adic realization M_p. Let $M^\vee$ be the dual motive. In the p-adic realization it corresponds to the dual Galois module. We denote by $H^1_{\mathcal{M}}(\mathbb{Z}, M(k))$ the "integral" motivic cohomology of the motive M in the sense of Beilinson [1].

For any finite Galois extension $K/\mathbb{Q}$ with Galois group G, let $\mathbb{Q}[G]$ be the group ring of G. It is a non-commutative ring with center denoted $Z(\mathbb{Q}[G])$.

We consider the *deformation* $\mathbb{Q}[G] \otimes M := h^0(K) \otimes M$. If $M = h^r(X)$ and $K/\mathbb{Q}$ is a number field, then $h^0(K) \otimes M = h^r(X \times K)$ considered as a motive over $\mathbb{Q}$.

We consider a finite set of primes S satisfying:

(*) $\mathbb{Q}[G] \otimes M$ and K have good reduction at all primes not dividing S, and $p \in S$.

2.2. Equivariant L-functions

We *assume* the usual conjectures about the L-functions of motives, like meromorphic continuation and functional equation etc., to be satisfied.

In order to define the *equivariant* L-function for G (without the Euler factors at the primes dividing S), consider a Galois extension $E/\mathbb{Q}$ such that $E[G] \cong \bigoplus_\rho \mathrm{End}_E(V(\rho))$, where $V(\rho)$ are absolutely irreducible representations of G. Then the center of $E[G]$ is $Z(E[G]) \cong \bigoplus_\rho E$ and the motives $V(\rho) \otimes M$ have coefficients in E. We define

$$L_S(G, M, k)^* := \left(L_S(V(\rho) \otimes M, k)^*\right)_\rho \in Z(E \otimes_{\mathbb{Q}} \mathbb{C}[G])^*$$

to be the element with ρ-component the leading coefficient at $s = k$ of the $E \otimes_{\mathbb{Q}} \mathbb{C}$-valued L-functions $L_S(V(\rho) \otimes M, s)$ without the Euler factors at S. Then $L_S(G, M, k)^*$ has actually values in $Z(\mathbb{R}[G])^*$ (see [4] Lemma 7) and is independent of the choice of E. We will always consider $L_S(G, M, k)^*$ as an element in $Z(\mathbb{R}[G]) \subset \mathbb{R}[G]$.

Remark In [22] Kato uses a different description of this equivariant L-function.

2.3. Non-commutative determinants

We follow the point of view of Burns and Flach. Let A be a (possibly non-commutative) ring and $V(A)$ the category of virtual objects in the sense of Deligne [12]. $V(A)$ is a monoidal tensor category and has a unit object $\mathbf{1}_A$. Moreover it is a groupoid, i.e., all morphisms are isomorphisms. There is a functor

$$\det_A : \{\text{perfect complexes of } A\text{-modules and isomorphisms}\} \to V(A)$$

which is multiplicative on short exact sequences. The group of isomorphism classes of objects of $V(A)$ is $K_0(A)$ and

$$\mathrm{Aut}(\mathbf{1}_A) = K_1(A) = \mathrm{Gl}_\infty(A)/E(A)$$

($E(A)$ the elementary matrices). In general $\mathrm{Hom}_{V(A)}(\det_A X, \det_A Y)$ is either empty or a $K_1(A)$-torsor. If $A \to B$ is a ring homomorphism, we get a functor $B\otimes : V(A) \to V(B)$ such that tensor product and $\det_A$ commute.

Convention By abuse of notation we are going to write $z \in \det_A X$ for $z : \mathbf{1}_A \to \det_A X$ and call this a *generator* of $\det_A X$.

If A is commutative and local, then the category of virtual objects is equivalent to the category of pairs (L, r) where L is an invertible A-module and $r \in \mathbb{Z}$. One recovers the theory of determinants of Knudson and Mumford. The unit object is $\mathbf{1}_A = (A, 0)$ and one has $\mathrm{Aut}(\mathbf{1}_A) = K_1(A) = A^*$. Thus $K_1(A)$ is used as generalization of A^* to the non-commutative case. Generators of $\det_A X = (L, 0)$ in the above sense correspond to A-generators of L.

2.4. Formulation of the conjecture

The original conjecture dates back to Beilinson [1] and Bloch-Kato [3]. The idea of an equivariant formulation is due to Kato [23] and [22]. Fontaine and Perrin-Riou gave a uniform formulation for mixed motives and all values of L-functions at

all integer values [14], [15]. The generalization to non-abelian coefficients is due to Burns and Flach [4].

For simplicity of exposition, we restrict to values at very negative integers. In the absolute case this coincides with the formulation given by Kato in [23]. We consider a motive M and values at $1-k$ where k is *big enough.* In the case $M = h^r(X)$, k big enough means that

- $k > \inf\{r, \dim(X)\}$, $(r,k) \neq (1,0); (2\dim(X), \dim(X)+1)$ and $2k \neq r+1$.
- for all $\ell \in S$ the local Euler factor $L_\ell(M_p^\vee, s)^{-1}$ at ℓ does not vanish at $1-k$.

Consider the (injective) reduced norm map $rn : K_1(\mathbb{R}[G]) \to Z(\mathbb{R}[G])^*$ and recall that $L_S(G, M^\vee, 1-k)^* \in Z(\mathbb{R}[G])^*$. By strong approximation (see [4] Lemma 8) there is $\lambda \in Z(\mathbb{Q}[G])^*$ such that $\lambda L_S(G, M^\vee(1-k))^*$ is in the image of $K_1(\mathbb{R}[G])$ under rn. Let

$$\lambda L_S(G, M^\vee(1-k))^* \in \mathbf{1}_{\mathbb{R}[G_n]}$$

be the corresponding generator. For k big enough, we define the *fundamental line* in $V(\mathbb{Q}[G])$ as

$$\Delta_f(G, M^\vee(1-k)) = \det^{-1}_{\mathbb{Q}[G]} H^1_{\mathcal{M}}(\mathbb{Z}, \mathbb{Q}[G] \otimes M(k)) \otimes \det_{\mathbb{Q}[G]}(\mathbb{Q}[G] \otimes M_B(k-1))^+ .$$

Here $+$ denotes the fixed part under complex conjugation.

Conjecture 2.4.1 *Let M be as in 2., $p \neq 2$ a prime and k big enough.*

1. *The Beilinson regulator $r_{\mathcal{D}}$ induces an isomorphism*

$$\Delta_f(G, M^\vee(1-k)) \otimes \mathbb{R} \cong \mathbf{1}_{\mathbb{R}[G]}.$$

2. *Under this isomorphism the generator $(\lambda L_S(G, M^\vee(1-k))^*)^{-1}$ is induced by a (unique) generator*

$$(\lambda^{-1}\delta(G, M, k)) \in \Delta_f(G, M^\vee(1-k)).$$

The reduced norm is an isomorphism $K_1(\mathbb{Q}_p[G]) \cong Z(\mathbb{Q}_p[G])^*$. Using the operation of $K_1(\mathbb{Q}_p[G])$ on generators in $\Delta_f(G, M^\vee(1-k)) \otimes \mathbb{Q}_p$, we put

$$\delta_p(G, M, k) := (\lambda^{-1}\delta(G, M, k))\lambda \in \Delta_f(G, M^\vee(1-k)) \otimes \mathbb{Q}_p.$$

Note that this generator is independent of the choice of λ.

3. *The p-adic regulator r_p induces an isomorphism*

$$\Delta_f(G, M^\vee(1-k)) \otimes \mathbb{Q}_p \cong$$
$$\det^{-1}_{\mathbb{Q}_p[G]} H^1(\mathbb{Z}[1/S], \mathbb{Q}_p[G] \otimes M_p(k)) \otimes \det_{\mathbb{Q}_p[G]}(\mathbb{Q}_p[G] \otimes M_B(k-1))^+.$$

4. *Let $T_B \subset M_B$ be a lattice such that $T_p = T_B \otimes \mathbb{Z}_p \subset M_p$ is Galois stable. Under the last isomorphism $\delta_p(G, M, k)$ is induced by a generator*

$$\widetilde{\delta}_p(G, M, k) \in \det_{\mathbb{Z}_p[G]} R\Gamma(\mathbb{Z}[1/S], \mathbb{Z}_p[G] \otimes T_p(k)) \otimes \det_{\mathbb{Z}_p[G]}(\mathbb{Z}_p[G] \otimes T_B)(k-1))^+.$$

Remark a) The conjecture is compatible with change of group G. If $G \to G'$ is a surjection, then the equivariant conjecture for G tensored with $\mathbb{Q}[G']$ over $\mathbb{Q}[G]$ gives the conjecture for G'.
b) The element $\widetilde{\delta}_p(G, M, k)$ is determined up to an element in the kernel of the map $K_1(\mathbb{Z}_p[G]) \to K_1(\mathbb{Q}_p[G])$. In the commutative case, this map is always injective.
c) The conjecture is independent of T. It is also independent of S. This computation shows that the definition of the equivariant L-function forces the use of the reduced norm in the formulation of the conjecture.

3. Non-abelian Main Conjecture

3.1. Iwasawa algebra and modules

Let K_n be a tower of finite Galois extensions of $\mathbb{Q}$ with Galois groups G_n such that $G_\infty := \varprojlim G_n$ is a p-adic Lie group of dimension at least 1. Moreover, we assume that only finitely many primes ramify in $K_\infty := \bigcup_n K_n$.

The classical example is the cyclotomic tower $K_n := \mathbb{Q}(\zeta_{p^n})$ with ζ_{p^n} a p^n-th root of unity. A non-abelian example is the tower $K_n := \mathbb{Q}(E[p^n])$, where $E[p^n]$ are the p^n-torsion points of an elliptic curve E without CM defined over $\mathbb{Q}$.

The *Iwasawa algebra*

$$\Lambda := \mathbb{Z}_p[[G_\infty]] = \varprojlim \mathbb{Z}_p[G_n]$$

is the ring of $\mathbb{Z}_p$-valued distributions on G_∞. It is a possibly non-commutative Noetherian semi-local ring. If G_∞ is in addition a pro-p-group without p-torsion it is even a regular and local ring.

For the cyclotomic tower, $\Lambda \cong \mathbb{Z}_p[G_1][[t]]$ is the classical Iwasawa algebra. For the tower of p^n-torsion points of E, the Iwasawa algebra was studied by Coates and Howson [8], [9]. Modules over such algebras are studied recently by Venjakob [36] and by Coates-Schneider-Sujatha [10].

We are concerned with the complex of Λ-modules $R\Gamma(\mathbb{Z}[1/S], \Lambda \otimes_{\mathbb{Z}_p} T_p(k))$ and $(\Lambda \otimes_{\mathbb{Z}} T_B(k-1))^+$. They are perfect complexes. Note that

$$R\Gamma(\mathbb{Z}[1/S], \Lambda \otimes_{\mathbb{Z}_p} T_p(k)) = \varprojlim R\Gamma(\mathcal{O}_{K_n}[1/S], T_p(k))$$

where $\mathcal{O}_{K_n}$ is the ring of integers of K_n.

3.2. Formulation of the non-abelian Main Conjecture

The Main Conjecture can be viewed as a Bloch-Kato conjecture for the deformed "motive" $\Lambda \otimes M$ with coefficients in Λ.

Recall from 2.4.1 that the generators $\delta_p(G_n, M, k)$ are compatible under the transition maps $\mathbb{Q}_p[G_n] \to \mathbb{Q}_p[G_{n-1}]$. They define

$$\delta_p(G_\infty, M, k) = \varprojlim \delta_p(G_n, M, k) \in$$
$$\varprojlim \left[\det_{\mathbb{Q}_p[G_n]} R\Gamma(\mathbb{Z}[1/S], \mathbb{Q}_p[G_n] \otimes M_p(k)) \otimes \det_{\mathbb{Q}_p[G_n]} (\mathbb{Q}_p[G_n] \otimes M_B(k-1))^+ \right],$$

more precisely an element of $\varprojlim \mathrm{Hom}_{V(\mathbb{Q}_p[G_n])}(\mathbf{1}_{\mathbb{Q}_p[G_n]}, \cdot\,)$.

The map $\Lambda \to \mathbb{Q}_p[G_n]$ induces an isomorphism

$$\mathbb{Q}_p[G_n] \otimes_\Lambda R\Gamma\big(\mathbb{Z}[1/S], \Lambda \otimes T_p(k)\big) \to R\Gamma\big(\mathbb{Z}[1/S], \mathbb{Q}_p[G_n] \otimes_{\mathbb{Q}_p} M_p(k)\big)\ .$$

Conjecture 3.2.1 (Non-abelian Main Conjecture) *Let M and S be as in 2., G_∞ as in 3., $p \neq 2$, $T_B \subset M_B$ a lattice such that $T_p := T_B \otimes \mathbb{Z}_p$ is Galois stable and k big enough (cf. section 2.). Then $\delta_p(G_\infty, M, k)$ is induced by a generator*

$$\widetilde{\delta}_p(G_\infty, M, k) \in \left[\det_\Lambda R\Gamma(\mathbb{Z}[1/S], \Lambda \otimes T_p(k)) \otimes \det_\Lambda(\Lambda \otimes T_B(k-1))^+\right]\ .$$

The conjecture translates into the Iwasawa Main Conjecture in the case of Dirichlet characters or CM-elliptic curves. See section 5. for more details.

Remark a) The conjecture is independent of the choice of lattice T_B. The correction factor $(\Lambda \otimes T_B(k-1))^+$ compensates different choices of lattice.
b) Perrin-Riou [31] has defined a p-adic L-function and stated a Main Conjecture for motives in the abelian case. She starts at the other side of the functional equation, where the exponential map of Bloch-Kato comes into play. Her main tool is the "logarithme élargi", which maps Galois cohomology over K_∞ to a module of p-adic analytic nature. It would be interesting to compare her approach with the above.
c) A Main Conjecture for motives and the cyclotomic tower was formulated by Greenberg [16], [17]. Ritter and Weiss consider the case of the cyclotomic tower over a finite non-abelian extension [32].

Proposition 3.2.2 (see section 6.) *The equivariant Bloch-Kato conjecture for M, k and all G_n is equivalent to the Main Conjecture for M, k and G_∞.*

3.3. Twisting

Assume that T_p becomes trivial over K_∞, for example let G_∞ be the image of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ in $\mathrm{Aut}(T_p)$. Let T_p^{triv} be the $\mathbb{Z}_p$-module underlying T_p with trivial operation of the Galois group. The map $g \otimes t \mapsto g \otimes g^{-1}t$ induces an equivariant isomorphism $\Lambda \otimes_{\mathbb{Z}_p} T_p \cong \Lambda \otimes_{\mathbb{Z}_p} T_p^{\mathrm{triv}}$. Hence there is an isomorphism

$$\begin{aligned}\det_\Lambda R\Gamma(\mathbb{Z}[1/S], \Lambda \otimes T_p(k)) \otimes \det_\Lambda(\Lambda \otimes T_B(k-1))^+ \cong \\ \det_\Lambda R\Gamma(\mathbb{Z}[1/S], \Lambda \otimes T_p^{\mathrm{triv}}(k)) \otimes \det_\Lambda(\Lambda \otimes T_B^{\mathrm{triv}}(k-1))^+.\end{aligned}$$

Note that T_B^{triv} can be viewed as a lattice in the Betti-realization of the trivial motive $h^0(\mathbb{Q}) \otimes M^{\mathrm{triv}} = \mathbb{Q}(0) \otimes M^{\mathrm{triv}}$ where M^{triv} is M_B considered as $\mathbb{Q}$-vector space.

Corollary 3.3.1 *If the Main Conjecture is true for M and $\mathbb{Q}(0) \otimes M^{\mathrm{triv}}$ and k, then*

$$\widetilde{\delta}_p(G_\infty, M, k) = \widetilde{\delta}_p(G_\infty, M^{\mathrm{triv}}, k)$$

up to an element in $K_1(\Lambda)$ under the above isomorphism.

Remark Even if T_p is not trivial over K_∞, the same method allows to twist with a motive whose p-adic realization is trivial over K_∞. A particular interesting case is the motive $\mathbb{Q}(1)$ if K_∞ contains the cyclotomic tower. It allows to pass from values of the L-function at k to values at $k+1$.

Strategy This observation allows the following strategy for proving the Main Conjecture and the Bloch-Kato conjecture for all motives:

- first prove the equivariant Bloch-Kato conjecture for the motive $h^0(\mathbb{Q}) = \mathbb{Q}(0)$, one fixed k and all finite groups G_n. For $k = 1$ this is an equivariant class number formula.
- by proposition 3.2.2 this implies the Main Conjecture for the motives $\mathbb{Q}(k) \otimes M^{\mathrm{triv}}$ and all p-adic Lie groups G_∞.
- for any motive M there is a K_∞ such that T_p becomes trivial. Using corollary 3.3.1 it remains to show that $\widetilde{\delta}_p(G_\infty, M^{\mathrm{triv}}, k)$ induces $\delta_p(G_n, M, k)$ for all n. This is a compatibility conjecture for elements in motivic cohomology and allows to reduce to the case of number fields.
- the equivariant Bloch-Kato conjecture follows by 3.2.2.

4. Relation to classical Iwasawa theory in the critical case

4.1. Characteristic ideals

We restrict to the case G_∞ a pro-p-group without p-torsion. In this case the Iwasawa algebra is local and Auslander regular ([36]). Its total ring of quotients is a skew field D. Then $K_0(\Lambda) \cong K_0(D) \cong \mathbb{Z}$, $K_1(\Lambda) = (\Lambda^*)^{\mathrm{ab}}$, and $K_1(D) = (D^*)^{\mathrm{ab}}$ where $\cdot^{\mathrm{ab}}$ denotes the abelianization of the multiplicative group.

Let $\mathcal{T}$ be the category of finitely generated Λ-torsion modules. The localization sequence for K-groups implies an exact sequence

$$(\Lambda^*)^{\mathrm{ab}} \to (D^*)^{\mathrm{ab}} \to K_0(\mathcal{T}) \to 0.$$

If X is a Λ-torsion module, then we call its class in $K_0(\mathcal{T})$ the *characteristic ideal.* By the above sequence it is an element of D^* up to $[D^*, D^*] \operatorname{Im} \Lambda^*$. If G_∞ is abelian, $K_0(\mathcal{T})$ is nothing but the group of fractional ideals that appears in classical Iwasawa theory.

The characteristic ideal can also be computed from the theory of determinants. The class of X in $K_0(\Lambda)$ is necessarily 0, hence there exists a generator $x \in \det_\Lambda(X)$. Its image in $D \otimes \det_\Lambda(X) = \det_D(0) = \mathbf{1}_D$ is an element of $K_1(D)$. This construction yields a well-defined element of $K_1(D)/\operatorname{Im} K_1(\Lambda) \cong K_0(\mathcal{T})$, in fact the inverse of the characteristic ideal of X.

Note that a complex is perfect if and only if it is a bounded complex with finitely generated cohomology. Such complexes also have characteristic ideals if

their cohomology is Λ-torsion.

Remark Coates, Schneider and Sujatha study the category of Λ-torsion modules in [10]. In particular, they also define a notion of characteristic ideal as object of $K_0(\mathcal{T}^b/\mathcal{T}^1)$ where $\mathcal{T}^b/\mathcal{T}^1$ denotes the quotient category of bounded finitely generated Λ-torsion modules by the sub-category of pseudo-null modules. They construct a map

$$K_0(\mathcal{T}) \to K_0(\mathcal{T}/\mathcal{T}^1) \to K_0(\mathcal{T}^b/\mathcal{T}^1)$$

which maps the class of a module to the characteristic ideal in their sense. If G_∞ is abelian, then the two maps are isomorphisms and all notions of characteristic ideals agree. In the general case, we do not know whether the map is injective. However, it seems to us that the problem is not so much in passing to the quotient category modulo pseudo-null modules but rather in projecting to the bounded part.

4.2. Zeta distributions

Let M, k, S and G_∞ as before. Assume

$$H^1_{\mathcal{M}}(\mathbb{Z}, \mathbb{Q}[G_n] \otimes M(k)) = 0 \text{ for all } G_n.$$

For k big enough, this implies that $M_B(k-1)^+ = 0$ and K_n totally real. The motives $\mathbb{Q}[G_n] \otimes M(k)$ are *critical* in the sense of Deligne. Note that the only motives expected to be critical and to satisfy our condition k big enough (see 2.) are Artin motives (with $k > 1$).

In this case, the Beilinson conjecture asserts that $L_S(G_n, M^\vee, 1-k) \in Z(\mathbb{Q}[G_n])^*$ (no leading coefficients has to be taken). We call

$$\mathcal{L}_S(G_\infty, M^\vee, 1-k) = \varprojlim L_S(G_n, M^\vee, 1-k) \in \varprojlim Z(\mathbb{Q}_p[G_n])^*$$

the *zeta distribution.*

Let $f, g \in \Lambda$ such that the images $f_n, g_n \in \mathbb{Z}_p[G_n]$ are units in $\mathbb{Q}_p[G_n]$. Via the reduced norm, they define a distribution

$$(\mathrm{rn}(f_n g_n^{-1}))_n \in \varprojlim Z(\mathbb{Q}_p[G_n])^*.$$

Remark It is not clear to us if the class of $f/g \in K_1(D) = (D^*)^{\mathrm{ab}}$ is uniquely determined by the sequence $f_n g_n^{-1}$. In the abelian case this is true and f/g is a generalization of Serre's pseudo measure (cf. [35]).

In this case the complexes $R\Gamma(\mathcal{O}_{K_n}[1/S], T_p(k))$ are torsion. Hence the complex $R\Gamma(\mathbb{Z}[1/S], \Lambda \otimes T_p(k)) = \varprojlim_n R\Gamma(\mathcal{O}_{K_n}[1/S], T_p(k))$ is bounded and its cohomology is Λ-torsion (see [18]). The main conjecture 3.2.1 takes the following form:

Conjecture 4.2.1 *Let M be an Artin motive, $k > 1$, S, G_∞ as before (in particular G_∞ pro-p and without p-torsion) and $\mathbb{Q}[G_n] \otimes M(k)$ critical for all n. There exist*

$f, g \in \Lambda$ such that the induced distribution $(\mathrm{rn}(f_n g_n^{-1}))_n \in \varprojlim Z(\mathbb{Q}_p[G_n])^*$ is the zeta distribution $\mathcal{L}_S(G_\infty, M^\vee, 1-k)$ and the characteristic ideal

$$[R\Gamma(\mathbb{Z}[1/S], \Lambda \otimes T_p(k))[1]] \in K_0(\mathcal{T})$$

coincides with the image of $fg^{-1} \in (D^*)^{\mathrm{ab}}$.

Remark a) The conjecture is isogeny invariant, i.e., independent of the choice of lattice T_p. The correction term $(\Lambda \otimes T_B(k))^+$ vanishes.
b) In the abelian case this means that the zeta distribution is a pseudo measure and generates the characteristic ideal.
c) In the case of the cyclotomic tower, a similar conjecture is formulated by Greenberg, [16], [17].
d) If G_∞ is abelian, the above conjecture is easily seen to be implied by conjecture 3.2.1. The argument also works in the non-abelian case if the set of all elements of Λ which, for all n, are units in $\mathbb{Q}_p[G_n]$ is an Ore set.

5. Examples

5.1. Dirichlet characters

Let χ be a Dirichlet character, $V(\chi)$ its associated motive with coefficients in E. Let $\mathbb{Q}_\infty = \bigcup_n \mathbb{Q}_n$ be the cyclotomic $\mathbb{Z}_p$-extension of $\mathbb{Q}$ and $G_\infty = \mathrm{Gal}(\mathbb{Q}_\infty/\mathbb{Q}) = \varprojlim_n G_n$. In this case the equivariant L-function is $L_S(G_n, V(\chi), s) = (L_S(\rho\chi, s))_\rho$, where ρ runs through all characters of G_n and $L_S(\rho\chi, s)$ is the Dirichlet L-function associated to $\rho\chi$. Let k be big enough, i.e., $k > 1$.
Critical case $\chi(-1) = (-1)^k$.

Here $H^1_{\mathcal{M}}(\mathbb{Z}, E[G_n] \otimes V(\chi)(k)) = 0$ for all n. As in section 4., the equivariant L-values give rise to the zeta distribution $\mathcal{L}_S(G_\infty, V(\chi)^\vee, 1-k) \in \varprojlim_n E[G_n]$. It is a classical calculation (Stickelberger elements) that this is in fact a pseudo measure, which gives rise to the Kubota-Leopoldt p-adic L-function. Let $\mathcal{O} \subset E$ be the ring of integers, $\Lambda = \mathcal{O}_p[[G_\infty]]$ the Iwasawa algebra and $T_p(\chi) \subset V_p(\chi)$ a Galois stable lattice. The Iwasawa Main Conjecture 4.2.1 amounts to the following theorem:

Theorem 5.1.1 *The zeta distribution $\mathcal{L}_S(G_\infty, V(\chi)^\vee, 1-k)$ generates*

$$\det_\Lambda^{-1} H^1(\mathbb{Z}[1/S], \Lambda \otimes T_p(\chi)(k)) \otimes \det_\Lambda H^2(\mathbb{Z}[1/S], \Lambda \otimes T_p(\chi)(k)).$$

Remark This is a reformulation of the main theorem of Mazur and Wiles in [29]. There is an extension to the case of totally real fields by Wiles [37] and an equivariant version by Burns and Greither [6].
Non-critical case $\chi(-1) = (-1)^{k-1}$.

Here $H^1_{\mathcal{M}}(\mathbb{Z}, E[G_n] \otimes V(\chi)(k))$ has $E[G_n]$-rank 1. It is a theorem of Borel (resp. Soulé) that $r_{\mathcal{D}} \otimes \mathbb{R}$ (resp. $r_p \otimes \mathbb{Q}_p$) is an isomorphism. By a theorem of Beilinson-Deligne (see [21] or [19]), the image of $\delta_p(G_n, V(\chi), k)$ under r_p is given by

$$c_k(G_n, t_p(\chi))^{-1} \otimes t_p(\chi)(k-1),$$

where

$$c_k(G_n, t_p(\chi)) \in H^1(\mathbb{Z}[1/S], \mathcal{O}_p[G_n] \otimes T_p(\chi)(k))$$

is a twist of a cyclotomic unit and $t_p(\chi)(k-1)$ is a generator of $T_p(\chi)(k-1)$. Let $c_k(G_\infty, t_p(\chi)) := \varprojlim_n c_k(G_n, t_p(\chi))$.

Theorem 5.1.2 *There is a canonical isomorphism of Λ-determinants*

$$\det_\Lambda \left(H^1(\mathbb{Z}[1/S], \Lambda \otimes T_p(\chi)(k))/c_k(G_\infty, t_p(\chi))\right) \cong \det_\Lambda H^2(\mathbb{Z}[1/S], \Lambda \otimes T_p(\chi)(k)).$$

Remark For $p \nmid \mathrm{ord}(\chi)$ this is a consequence of theorem 5.1.1 and was shown directly by Rubin [33] with Euler system methods. The restriction at the order of χ is removed in Burns-Greither [5] and Huber-Kings [20] by different methods.

The Tamagawa number conjecture for $V(\chi)(r)$ (and hence for $h^0(F)(r)$ with F an abelian number field) can be deduced from theorems 5.1.1 and 5.1.2, see Burns-Greither [5] or Huber-Kings [20]. Previous partial results were proved in Mazur-Wiles [29], Wiles [37], Kato [22], [23], Kolster-Nguyen Quang Do-Fleckinger [26] and Benois-Nguyen Quang Do[2].

We would like to stress that the strategy 3. is used in Huber-Kings [20] to prove theorems 5.1.1, 5.1.2 and the Tamagawa number conjecture from the class number formula.

5.2. Elliptic curves

Let E be an elliptic curve over an imaginary quadratic field K with CM by $\mathcal{O}_K$. The motive $h^1(E)$ considered with coefficients in K decomposes into $V(\psi) \oplus V(\bar{\psi})$, where ψ is the Grössencharacter associated to E. The L-function of $V(\psi)$ is the Hecke L-function of ψ, which has a zero of order 1 at $2-k$, where $k \geq 2$. Let $S = Np$, where N is the conductor of ψ and let $K_n := K(E[p^n])$.

It is not known if $H^1_{\mathcal{M}}(\mathcal{O}_K, K[G_n] \otimes V(\psi)(k))$ has $K[G_n]$-rank 1 but Deninger [13] shows that $r_{\mathcal{D}} \otimes \mathbb{R}$ is surjective and that the Beilinson conjecture holds. It is a result of Kings [25] that the image in étale cohomology of the zeta element $\delta_p(G_n, V(\psi), 2-k)$ given by Beilinson's Eisenstein symbol is given by

$$e_k(G_n, t_p(\psi))^{-1} \otimes t_p(\psi) \ ,$$

where $e_k(G_n, t_p(\psi)) \in H^1(\mathbb{Z}[1/S], \mathcal{O}_p[G_n] \otimes T_p(\psi)(k))$ is the twist of an elliptic unit. Let $\Lambda := \mathcal{O}_p[[G_\infty]]$ and $e_k(G_\infty, t_p(\psi)) = \varprojlim_n e_k(G_n, t_p(\psi))$.

Theorem 5.2.1 *There is a canonical isomorphism of determinants*

$$\det_\Lambda \left(H^1(\mathbb{Z}[1/S], \Lambda \otimes T_p(\psi)(k))/e_k(G_\infty, t_p(\psi))\right) \cong \det_\Lambda H^2(\mathbb{Z}[1/S], \Lambda \otimes T_p(\psi)(k)).$$

Remark 1) This is a reformulation of Rubin's Iwasawa Main Conjecture [34].
2) In [25] the (absolute) Bloch-Kato conjecture for $V(\psi)$ is deduced from this under

the condition that $H^2(\mathbb{Z}[1/S], T_p(\psi)(k))$ is finite (fulfilled for almost all k for fixed p).

Kato [24] has investigated the case of elliptic curves over $\mathbb{Q}$ and the cyclotomic tower. His approach to the Birch-Swinnerton-Dyer conjecture uses the idea of twisting cup-products of Eisenstein symbols to the value of the L-function at 1. As a consequence he can prove one inclusion of the Iwasawa main conjecture in this case. The result supports our general philosophy of twisting to the case of number fields.

6. Proof of proposition 3.2.2

We want to give the proof of proposition 3.2.2. The implication from the Main Conjecture to the equivariant Bloch-Kato conjecture is trivial. Conversely, we have to show the following abstract statement:

Lemma 6.1 *Let $\nabla \in V(\Lambda)$ and $\widetilde{\delta}(n) \in \mathbb{Z}_p[G_n] \otimes \nabla$ generators such that their images $\delta(n) \in \mathbb{Q}_p[G_n] \otimes \nabla$ are compatible under transition maps. Then there is a generator $\widetilde{\delta}'(\infty) \in \nabla$ inducing all $\delta(n)$.*

The proposition follows with $\widetilde{\delta}(n) = \widetilde{\delta}_p(G_n, M, k)$ and

$$\nabla = \det_\Lambda R\Gamma(\mathbb{Z}[1/pS], \Lambda \otimes T_p(k)) \otimes \det_\Lambda(\Lambda \otimes T_B(k-1))^+.$$

We now prove the lemma. We first reduce to a statement about elements of K_1. By assumption, $\mathbb{Z}_p[G_n] \otimes \nabla$ has a generator, in particular, its isomorphism class is zero in $K_0(\mathbb{Z}_p[G_n])$. As $K_0(\Lambda) \to \varprojlim K_0(\mathbb{Z}_p[G_n])$ is an isomorphism, this implies that the class of ∇ is zero in $K_0(\Lambda)$. Without loss of generality we can assume $\nabla = \mathbf{1}_\Lambda$. Recall that by our convention, a generator of $\mathbf{1}_A$ is nothing but an element of the abelian group $K_1(A)$ for all rings A.

Let $B_n = \operatorname{Im} K_1(\mathbb{Z}_p[G_n]) \to K_1(\mathbb{Q}_p[G_n])$. By assumption $\delta(n) \in B_n$. There is a system of short exact sequences

$$0 \to SK_1(\mathbb{Z}_p[G_n]) \to K_1(\mathbb{Z}_p[G_n]) \to B_n \to 0.$$

By [11] 45.22 the groups $SK_1(\mathbb{Z}_p[G_n])$ are finite. The system of these groups is automatically Mittag-Leffler. Hence we get a surjective map

$$\varprojlim K_1(\mathbb{Z}_p[G_n]) \to \varprojlim B_n.$$

The system $(\delta(n))_n$ has a preimage $(\widetilde{\delta}'(n))_n \in \varprojlim K_1(\mathbb{Z}_p[G_n])$.

All $\mathbb{Z}_p[G_n]$ are semi-local, hence by [11] 40.44

$$K_1(\mathbb{Z}_p[G_n]) \cong \mathrm{Gl}_2(\mathbb{Z}_p[G_n])/E_2(\mathbb{Z}_p[G_n])$$

where E_2 is the subgroup of elementary matrices. We represent $\widetilde{\delta}'_p(n)$ by an element of $\mathrm{Gl}_2(\mathbb{Z}_p[G_n])$. By assumption the image of $\widetilde{\delta}'(n)$ in $K_1(\mathbb{Z}_p[G_{n-1}])$ differs from $\widetilde{\delta}'(n-1)$ by some elementary matrix in $E_2(\mathbb{Z}_p[G_{n-1}])$. Elementary matrices can be lifted to elementary matrices in $\mathrm{Gl}_2(\mathbb{Z}_p[G_n])$. Hence we can assume that the

elements $\widetilde{\delta}'(n)) \in \mathrm{Gl}_2(\mathbb{Z}_p[G_n])$ form a projective system. The system defines an element

$$\widetilde{\delta}'_p(n) \in \mathrm{Gl}_2(\Lambda)$$

whose class in $K_1(\Lambda)$ has the necessary properties.

References

[1] A. Beilinson, Higher regulators and values of L-functions, *Jour. Soviet. Math.*, 30 (1985), 2036–2070.

[2] D. Benois, Thong Nguyen Quang Do, La conjecture de Bloch et Kato pour les motifs $\mathbb{Q}(m)$ sur un corps abélien, Preprint 2000.

[3] S. Bloch, K. Kato, L-functions and Tamagawa numbers of motives, *The Grothendieck Festschrift*, Vol. I, 333–400, Progr. Math., 86, Birkhäuser Boston, Boston, MA, 1990.

[4] D. Burns, M. Flach, Tamagawa numbers for motives with (non-commutative) coefficients, *Doc. Math.*, 6 (2001), 501–570 (electronic).

[5] D. Burns, C. Greither, On the equivariant Tamagawa conjecture for Tate motives, Preprint 2001.

[6] D. Burns, C. Greither, Equivariant Weierstrass preparation and values of L-functions at negative integers, preprint 2002.

[7] J. Coates, Fragments of the GL_2 Iwasawa theory of elliptic curves without complex multiplication. *Arithmetic theory of elliptic curves (Cetraro, 1997)*, 1–50, Lecture Notes in Math., 1716.

[8] J. Coates, S. Howson, Euler characteristics and elliptic curves, *Elliptic curves and modular forms (Washington, DC, 1996)*, Proc. Nat. Acad. Sci. U.S.A. 94 (1997), no. 21, 11115–11117.

[9] J. Coates, S. Howson, Euler characteristics and elliptic curves. II, *J. Math. Soc. Japan*, 53 (2001), no. 1, 175–235.

[10] J. Coates, P. Schneider, R. Sujatha, Modules over Iwasawa algebras, Preprint 2001.

[11] C.W. Curtis, I. Reiner, *Methods of representation theory*, Vol. I. and Vol. II, John Wiley & Sons, Inc., New York, 1981 and 1987.

[12] P. Deligne, Le déterminant de la cohomologie, *Current trends in arithmetical algebraic geometry (Arcata, Calif., 1985)*, 93–177, Contemp. Math., 67, Amer. Math. Soc., Providence, RI, 1987.

[13] C. Deninger, Higher regulators and Hecke L-series of imaginary quadratic fields I, *Invent. Math.*, 96 (1989), no. 1, 1–69.

[14] J.-M. Fontaine, Valeurs spéciales des fonctions L des motifs, *Séminaire Bourbaki*, Vol. 1991/92. Astérisque No. 206, (1992), Exp. No. 751, 4, 205–249.

[15] J.-M. Fontaine, B. Perrin-Riou, Autour des conjectures de Bloch et Kato: cohomologie galoisienne et valeurs de fonctions L, *Motives (Seattle, WA, 1991)*, 599–706, Proc. Sympos. Pure Math., 55, Part 1, Amer. Math. Soc., Providence, RI, 1994.

[16] R. Greenberg, Iwasawa theory for motives, *L-functions and arithmetic, Pro-*

ceedings of the Durham Symposium 1989, LMS Lecture Notes Series Vol. 153, Cambridge University press, 211–234.

[17] R. Greenberg, Iwasawa Theory and p-adic Deformation of Motives, *Motives (Seattle, WA, 1991)*, 193–223, Proc. Sympos. Pure Math., 55, Part 2, Amer. Math. Soc., Providence, RI, 1994.

[18] M. Harris, p-adic representations arising from descent on Abelian Varieties, Harvard PhD Thesis 1977, also *Comp. Math.*, 39 (1979), 177-245.

[19] A. Huber, G. Kings, Degeneration of l-adic Eisenstein classes and of the elliptic polylog, *Invent. math.*, 135 (1999), 545–594.

[20] A. Huber, G. Kings, Bloch-Kato Conjecture and Main Conjecture of Iwasawa Theory for Dirichlet Characters, Preprint 2001, revised 2002. To appear: *Duke Math. Journal.*

[21] A. Huber, J. Wildeshaus, Classical motivic polylogarithm according to Beilinson and Deligne, *Doc. Math. J. DMV*, 3 (1998), 27–133 and 297–299.

[22] K. Kato, Iwasawa theory and p-adic Hodge theory, *Kodai Math. J.*, 16 (1993), no. 1, 1–31.

[23] K. Kato, Lectures on the approach to Iwasawa theory for Hasse-Weil L-functions via B_{dR}. I, *Arithmetic algebraic geometry (Trento, 1991)*, 50–163, Lecture Notes in Math., 1553, Springer, Berlin, 1993.

[24] K. Kato, p-adic Hodge theory and values of zeta functions of modular forms, preprint 2000.

[25] G. Kings, The Tamagawa number conjecture for CM elliptic curves, *Invent. Math.*, 143 (2001), no. 3, 571–627.

[26] M. Kolster, Th. Nguyen Quang Do, V. Fleckinger, Twisted S-units, p-adic class number formulas, and the Lichtenbaum conjectures, *Duke Math. J.*, 84 (1996), no. 3, and Correction, *Duke Math. J.*, 90 (1997), no. 3, 641–643.

[27] M. Kolster, Th. Nguyen Quang Do, Universal distribution lattices for abelian number fields, Preprint 2000.

[28] T.Y. Lam, *A first course in noncommutative rings*, Second edition, Graduate Texts in Mathematics, 131.

[29] B. Mazur, A. Wiles, Class fields of abelian extensions of $\mathbb{Q}$, *Invent. Math.*, 76 (1984), no. 2, 179–330.

[30] Y. Ochi, O. Venjakob, On the structure of Selmer groups of p-adic Lie extensions, to appear: *Journal of Alg. Geom.*

[31] B. Perrin-Riou, *p-adic L-functions and p-adic representations*, SMF/AMS Texts and Monographs, 3., Paris, 2000.

[32] J. Ritter, A. Weiss, Toward equivariant Iwasawa theory, to appear: *Manuscripta Mathematica.*

[33] K. Rubin, *Euler Systems*, Annals of Mathematics Studies, 147, Princeton, NJ, 2000.

[34] K. Rubin, The "main conjectures" of Iwasawa theory for imaginary quadratic fields, *Invent. Math.*, 103 (1991), no. 1, 25–68.

[35] J-P. Serre: Sur le résidu de la fonction zêta p-adique d'un corps de nombres, *C.R. Acad. Sci. Paris*, Série A, 287 (1978), A183–A188.

[36] O. Venjakob, On the structure theory of the Iwasawa algebra of a p-adic Lie

group, to appear: *Jour. Eur. Math. Soc.*

[37] A. Wiles, The Iwasawa conjecture for totally real fields, *Annals of Math.*, 131 (1990), 493–540.

ICM 2002 · Vol. II · 163–171

Tamagawa Number Conjecture for zeta Values

Kazuya Kato*

Abstract

Spencer Bloch and the author formulated a general conjecture (Tamagawa number conjecture) on the relation between values of zeta functions of motives and arithmetic groups associated to motives. We discuss this conjecture, and describe some application of the philosophy of the conjecture to the study of elliptic curves.

2000 Mathematics Subject Classification: 11G40.
Keywords and Phrases: zeta function, Etale cohomology, Birch Swinnerton-Dyer conjecture.

Mysterious relations between zeta functions and various arithmetic groups have been important subjects in number theory.

(0.0) zeta functions $\leftrightarrow$ arithmetic groups.

A classical result on such relation is the class number formula discovered in 19th century, which relates zeta functions of number field to ideal class groups and unit groups. As indicated in (0.1)–(0.3) below, the formula of Grothendieck expressing the zeta functions of varieties over finite fields by etale cohomology groups, Iwasawa main conjecture proved by Mazur-Wiles, and Birch and Swinnerton-Dyer conjectures for abelian varieties over number fields, considered in 20th century, also have the form (0.0).

(0.1) Formula of Grothendieck.
zeta functions $\leftrightarrow$ etale cohomology groups.

(0.2) Iwasawa main conjecture.
zeta functions, zeta elements $\leftrightarrow$ ideal class groups, unit groups.

(0.3) Birch Swinnerton-Dyer conjectures (see 4).
zeta functions $\leftrightarrow$ groups of rational points, Tate-Shafarevich groups.

Here in (0.2), "zeta elements" mean cyclotomic units which are units in cyclotomic fields and closely related to zeta functions. Roughly speaking, the relations

*Department of Mathematical Sciences, University of Tokyo, Komaba 3-8-1, Meguro, Tokyo, Japan. E-mail: kkato@ms.u-tokyo.ac.jp

(often conjectural) say that the order of zero or pole of the zeta function at an integer point is equal to the rank of the related finitely generated arithmetic abelian group (Tate, the conjecture (0.3), Beilinson, Bloch, ...) and the value of the zeta function at an integer point is related to the order of the related arithmetic finite group.

In [BK], Bloch and the author formulated a general conjecture on (0.0) (Tamagawa number conjecture for motives). Further generalizations of Tamagawa number conjecture by Fontaine, Perrin-Riou, and the author [FP], [Pe_1] [Ka_1], [Ka_2] have the form

(0.4) zeta functions (= Euler products, analytic)
$\leftrightarrow$ zeta elements (= Euler systems, arithmetic)
$\leftrightarrow$ arithmetic groups.

Here the first $\leftrightarrow$ means that zeta functions enter the arithmetic world transforming themselves into zeta elements, and the second $\leftrightarrow$ means that zeta elements generate "determinants" of certain etale cohomology groups.

The aim of this paper is to discuss (0.4) in an expository style. We review (0.1) in §1, and then in §2, we describe the generalized Tamagawa number conjecture (0.4), the relation with (0.2), and an application of the philosophy (0.4) to (0.3).

In this paper, we fix a prime number p. For a commutative ring R, let $Q(R)$ be the total quotient ring of R obtained from R by inverting all non-zerodivisors.

1. Grothendieck formula and zeta elements

Let X be a scheme of finite type over a finite field $\mathbf{F}_q$. We assume p is different from $\mathrm{char}(\mathbf{F}_q)$.

In this §1, we first review the formula (1.1.2) of Grothendieck representing zeta functions of p-adic sheaves on X by etale cohomology. We then show that those zeta functions are recovered from p-adic zeta elements (1.3.5).

1.1. Zeta functions and etale cohomology groups in positive characteristic case. The Hasse zeta function $\zeta(X,s) = \prod_{x\in|X|}(1-\sharp\kappa(x)^{-s})^{-1}$, where $|X|$ denotes the set of all closed points of x and $\kappa(x)$ denotes the residue field of x, has the form $\zeta(X,s) = \zeta(X/\mathbf{F}_q, q^{-s})$ where

$$\zeta(X/\mathbf{F}_q,u) = \prod_{x\in|X|}(1-u^{\deg(x)})^{-1}, \qquad \deg(x) = [\kappa(x):\mathbf{F}_q]. \tag{1.1.1}$$

A part of Weil conjectures was that $\zeta(X/\mathbf{F}_q,u)$ is a rational function in u, and it was proved by Dwork and then slighly later by Grothendieck. The proof of Grothendieck gives a presentation of $\zeta(X/\mathbf{F}_q,u)$ by using etale cohomologyy. More generally, for a finite extension L of $\mathbf{Q}_p$ and for a constructible L-sheaf $\mathcal{F}$ on X, Grothendieck proved that the L-function $L(X/\mathbf{F}_q,\mathcal{F},u)$ has the presentation

$$L(X/\mathbf{F}_q,\mathcal{F},u) = \prod_m \det{}_L(1-\varphi_q u \ ; \ H^m_{et,c}(X\otimes_{\mathbf{F}_q}\bar{\mathbf{F}}_q,\mathcal{F}))^{(-1)^{m-1}} \tag{1.1.2}$$

where $H^m_{et,c}$ is the etale cohomology with compact supports and φ_q is the action of the q-th power morphism on X.

In the case $L = \mathbf{Q}_p = \mathcal{F}$, $\zeta(X/\mathbf{F}_q, u) = L(X/\mathbf{F}_q, \mathcal{F}, u)$.

1.2. p-adic zeta elements in positive characteristic case. Determinants appear in the theory of zeta functions as above, rather often. The regulator of a number field, which appears in the class number formula, is a determinant. Such relation with determinant is well expressed by the notion of "determinant module".

If R is a field, for an R-module V of dimension r, $\det_R(V)$ means the 1 dimensional R-module $\wedge^r_R(V)$. For a bounded complex C of R-modules whose cohomologies $H^m(C)$ are finite dimensional, $\det_R(C)$ means $\otimes_{m\in\mathbf{Z}} \{\det_R(H^m(C))\}^{\otimes(-1)^m}$.

This definition is generalized to the definition of an invertible R-module $\det_R(C)$ associated to a perfect complex C of R-modules for a commutative ring R (see [KM]). $\det^{-1}_R(C)$ means the inverse of the invertible module $\det_R(C)$.

By a pro-p ring, we mean a topological ring which is an inverse limit of finite rings whose orders are powers of p. Let Λ be a commutative pro-p ring. By a ctf Λ-complex on X, we mean a complex of Λ-sheaves on X for the etale topology with constructible cohomology sheaves and with perfect stalks. For a ctf Λ-complex $\mathcal{F}$ on X, $R\Gamma_{et,c}(X, \mathcal{F})$ ($_c$ means with compact supports) is a perfect complex over Λ.

For a commutative pro-p ring Λ and for a ctf Λ-complex $\mathcal{F}$ on X, we define the p-adic zeta element $\zeta(X, \mathcal{F}, \Lambda)$ which is a Λ-basis of $\det^{-1}_\Lambda R\Gamma_{et,c}(X, \mathcal{F})$. Consider the distinguished triangle

$$R\Gamma_{et,c}(X, \mathcal{F}) \to R\Gamma_{et,c}(X \otimes_{\mathbf{F}_q} \bar{\mathbf{F}}_q, \mathcal{F}) \xrightarrow{1-\varphi} R\Gamma_{et,c}(X \otimes_{\mathbf{F}_q} \bar{\mathbf{F}}_q, \mathcal{F}). \tag{1.2.1}$$

Since det is multiplicative for distinguished triangles, (1.2.1) induces an isomorphism

$$\det^{-1}_\Lambda R\Gamma_{et,c}(X, \mathcal{F}) \cong \det^{-1}_\Lambda R\Gamma_{et,c}(X \otimes_{\mathbf{F}_q} \bar{\mathbf{F}}_q, \mathcal{F}) \otimes_\Lambda \det_\Lambda R\Gamma_{et,c}(X \otimes_{\mathbf{F}_q} \bar{\mathbf{F}}_q, \mathcal{F}) \cong \Lambda. \tag{1.2.2}$$

We define $\zeta(X, \mathcal{F}, \Lambda)$ to be the image of $1 \in \Lambda$ in $\det^{-1}_\Lambda R\Gamma_{et,c}(X, \mathcal{F})$ under (1.2.2). It is a Λ-basis of the invertible Λ-module $\det^{-1}_\Lambda R\Gamma_{et,c}(X, \mathcal{F})$.

1.3. Zeta functions and p-adic zeta elements in positive characteristic case. Let L be a finite extension of $\mathbf{Q}_p$, let O_L be the valuation ring of L, and let $\mathcal{F}$ be a constructible O_L-sheaf on X. We show that the zeta function $L(X/\mathbf{F}_q, \mathcal{F}_L, u)$ of the L-sheaf $\mathcal{F}_L = \mathcal{F} \otimes_{O_L} L$ is recovered from a certain p-adic zeta element as in (1.3.5) below. Let

$$\Lambda = O_L[[\mathrm{Gal}(\bar{\mathbf{F}}_q/\mathbf{F}_q)]] = \varprojlim_n O_L[\mathrm{Gal}(\mathbf{F}_{q^n}/\mathbf{F}_q)]. \tag{1.3.1}$$

Let $s(\Lambda)$ be the Λ-module Λ which is regarded as a sheaf on the etale site of X via the natural action of $\mathrm{Gal}(\bar{\mathbf{F}}_q/\mathbf{F}_q)$. Then

$$H^m_{et,c}(X, \mathcal{F} \otimes_{O_L} s(\Lambda)) \cong \varprojlim_n H^m_{et,c}(X \otimes_{\mathbf{F}_q} \mathbf{F}_{q^n}, \mathcal{F}) \tag{1.3.2}$$

where the transition maps of the inverse system are the trace maps. From this, we can deduce that $H^m_{et,c}(X, \mathcal{F} \otimes_{O_L} s(\Lambda))$ is a finitely generated O_L-module for any m. Hence we have $Q(\Lambda) \otimes_\Lambda R\Gamma_{et,c}(X, \mathcal{F} \otimes_{O_L} s(\Lambda)) = 0$ and this gives an identificatition canonical isomorphism

$$Q(\Lambda) \otimes_\Lambda \det{}^{-1}_\Lambda R\Gamma_{et,c}(X, \mathcal{F} \otimes_{O_L} t(\Lambda)) = Q(\Lambda). \tag{1.3.3}$$

Note

$$Q(\Lambda) = Q(\varprojlim_n O_L[u]/(u^n - 1)) \supset Q(O_L[u]) = L(u). \tag{1.3.4}$$

By a formal argument, we can prove the following (1.3.5) (1.3.6) which show

zeta function = zeta element, zeta value = zeta element,

respectively.

$$L(X/\mathbf{F}_q, \mathcal{F}_L, u) = \zeta(X, \mathcal{F} \otimes_{O_L} s(\Lambda), \Lambda) \text{ in } Q(\Lambda). \tag{1.3.5}$$

If $H^m_{et,c}(X, \mathcal{F}_L) = 0$ for any m, $L(X/\mathbf{F}_q, \mathcal{F}_L, u)$ has no zero or pole at $u = 1$, and

$$L(X/\mathbf{F}_q, \mathcal{F}_L, 1) = \zeta(X, \mathcal{F}, O_L) \text{ in } L. \tag{1.3.6}$$

2. Tamagawa number conjecture

In 2.1, we describe the generalized version of Tamagawa number conjecture. In 2.2 (resp. 2.3), we consider p-adic zeta elements associated to 1 (resp. 2) dimensional p-adic representations of $\mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q})$, and their relations to (0.2) (resp. (0.3)).

2.1. The conjecture. Let X be a scheme of finite type over $\mathbf{Z}[\frac{1}{p}]$. For a complex of sheaves $\mathcal{F}$ on X for the etale topology, we define the compact support version $R\Gamma_{et,c}(X, \mathcal{F})$ of $R\Gamma_{et}(X, \mathcal{F})$ as the mapping fiber of

$$R\Gamma_{et}(\mathbf{Z}[\frac{1}{p}], Rf_!\mathcal{F}) \to R\Gamma_{et}(\mathbf{R}, Rf_!\mathcal{F}) \oplus R\Gamma_{et}(\mathbf{Q}_p, Rf_!\mathcal{F}).$$

where $f : X \to \mathrm{Spec}(\mathbf{Z}[\frac{1}{p}])$.

It can be shown that for a commutative pro-p ring Λ and for a ctf Λ-complex $\mathcal{F}$ on X, $R\Gamma_{et,c}(X, \mathcal{F})$ is perfect.

The following is a generalized version of the Tamagawa number conjecture [BK] (see [FP], [Pe$_1$], [Ka$_1$], [Ka$_2$]). In [BK], the idea of Tamagawa number of motives was important, but it does not appear explicitly in this version.

Conjecture. *To any triple $(X, \Lambda, \mathcal{F})$ consisting of a scheme X of finite type over $\mathbf{Z}[\frac{1}{p}]$, a commutative pro-p ring Λ, and a ctf Λ-complex on X, we can associate a Λ-basis $\zeta(X, \mathcal{F}, \Lambda)$ of*

$$\Delta(X, \mathcal{F}, \Lambda) = det^{-1}_\Lambda R\Gamma_{et,c}(X, \mathcal{F}),$$

which we call the p-adic zeta element associated to $\mathcal{F}$, satisfying the following conditions (2.1.1)-(2.1.5).
(2.1.1) If X is a scheme over a finite field $\mathbf{F}_q$, $\zeta(X,\mathcal{F},\Lambda)$ coincides with the element defined in §3.2.
(2.1.2) (rough form) If $\mathcal{F}$ is the p-adic realization of a motive M, $\zeta(X,\mathcal{F},\Lambda)$ recovers the complex value $\lim_{s\to 0} s^{-e}L(M,s)$ where $L(M,s)$ is the zeta function of M and e is the order of $L(M,s)$ at $s=0$.
(2.1.3) If Λ' is a pro-p ring and $\Lambda\to\Lambda'$ is a continuous homomorphism, $\zeta(X,\mathcal{F}\otimes^L_\Lambda \Lambda',\Lambda')$ coincides with the image of $\zeta(X,\mathcal{F},\Lambda)$ under $\Delta(X,\mathcal{F}\otimes^L_\Lambda\Lambda',\Lambda')\cong\Delta(X,\mathcal{F})\otimes_\Lambda \Lambda'$.
(2.1.4) For a distinguished triangle $\mathcal{F}'\to\mathcal{F}\to\mathcal{F}''$ with common X and Λ, we have

$$\zeta(X,\mathcal{F},\Lambda)=\zeta(X,\mathcal{F}',\Lambda)\otimes\zeta(X,\mathcal{F}'',\Lambda) \text{ in } \Delta(X,\mathcal{F},\Lambda)=\Delta(X,\mathcal{F}',\Lambda)\otimes_\Lambda\Delta(X,\mathcal{F}'',\Lambda).$$

(2.1.5) If Y is a scheme of finite type over $\mathbf{Z}[\frac{1}{p}]$ and $f: X\to Y$ is a separated morphism,

$$\zeta(Y,Rf_!\mathcal{F},\Lambda)=\zeta(X,\mathcal{F},\Lambda) \text{ in } \Delta(Y,Rf_!\mathcal{F},\Lambda)=\Delta(X,\mathcal{F},\Lambda).$$

By this (2.1.5), the constructions of p-adic zeta elements are reduced to the case $X=\mathrm{Spec}(\mathbf{Z}[\frac{1}{p}])$. How to formulate the part (4.1.2) of this conjecture is reduced to the case of motives over $\mathbf{Q}$ by (2.1.5) and $L(M,s)=L(Rf_!(M),s)$ (by philosophy of motives), where $f: X\to\mathrm{Spec}(\mathbf{Z}[\frac{1}{p}])$.

The conditions (2.1.3)-(2.1.5) are formal properties which are analogous to formal properties of zeta functions. The conditions (2.1.1) and (2.1.3)-(2.1.5) can be interpreted as

(2.1.6) The system $(X,\Lambda,\mathcal{F})\mapsto\zeta(X,\mathcal{F},\Lambda)$ is an "Euler system".

In fact, let L be a finite extension of $\mathbf{Q}_p$, S a finite set of prime numbers containing p, and let T be a free O_L-module of finite rank endowed with a continuous O_L-linear action of $\mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q})$ which is unramified outside S. For $m\geq 1$, let $R_m=O_L[\mathrm{Gal}(\mathbf{Q}(\zeta_m)/\mathbf{Q})]$ and let

$$z_m=\zeta_{R_m}(\mathbf{Z}[\tfrac{1}{p}],j_{m,!}(T\otimes_{O_L}s(R_m)),R_m)\in\det^{-1}_{R_m}R\Gamma_{et,c}(\mathbf{Z}[\zeta_m,\tfrac{1}{mS}],T).$$

($j_m:\mathrm{Spec}(\mathbf{Z}[\frac{1}{mS}])\to\mathrm{Spec}(\mathbf{Z}[\frac{1}{p}])$).

Then the conditions (4.1.1) and (4.1.3)-(4.1.5) tell that when m varies, the p-adic zeta elements z_m form a system satisfying the conditions of Euler systems formulated by Kolyvagin [Ko].

We illustrate the relation (2.1.2) with zeta functions.

Let M be a motive over $\mathbf{Q}$, that is, a direct summand of the motive $H^m(X)(r)$ for a proper smooth scheme X over $\mathbf{Q}$ and for $r\in\mathbf{Z}$, and assume that M is endowed with an action of a number field K. Then the zeta function $L(M,s)$ lives in $\mathbf{C}$, and the p-adic zeta element lives in the world of p-adic etale cohomology. Since these two worlds are too much different in nature, $L(M,s)$ and the p-adic zeta element are not simply related.

However in the middle of $\mathbf{C}$ and the p-adic world,

(a) there is a 1 dimensional K-vector space $\Delta_K(M)$ constructed by the Betti realization and the de Rham realization of M, and K-groups (or motivic cohomology groups) associated to M.

Let ∞ be an Archimedean place of K. Then

(b) there is an isomorphism

$$\Delta_K(M) \otimes_K K_\infty \underset{\cong}{\longrightarrow} K_\infty$$

constructed by Hodge theory and K-theory.

Let w be a place of K lying over p, let M_w be the representation of $\mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q})$ over K_w associated to M, and let T be a $\mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q})$-stable O_{K_w}-lattice in M_w. Then

(c) there is an isomorphism

$$\begin{aligned}\Delta_K(M) \otimes_K K_w &\underset{\cong}{\longrightarrow} \det{}^{-1}_{K_w} R\Gamma_{et,c}(\mathbf{Z}[\frac{1}{p}], j_* M_w) \\ =\det{}^{-1}_{O_{K_w}} R\Gamma_{et,c}(\mathbf{Z}[\frac{1}{p}], j_* T) &\otimes_{O_{K_w}} K_w\end{aligned}$$

where $j : \mathrm{Spec}(\mathbf{Q}) \to \mathrm{Spec}(\mathbf{Z}[\frac{1}{p}])$, constructed by p-adic Hodge theory and K-theory.

See [FP] how to construct (a)-(c) (constructions require some conjectures). The part (2.1.2) of the conjecture is:

(d) there exists a K-basis $\zeta(M)$ of $\Delta_K(M)$ (called the rational zeta element associated to M), which is sent to $\lim_{s\to 0} s^{-e} L(M,s)$ under the isomorphism (b) where e is the order of $L(M,s)$ at $s = 0$, and to $\zeta(\mathbf{Z}[\frac{1}{p}], j_* T, O_{K_w})$ in $\det^{-1}_{K_w} R\Gamma_{et,c}(\mathbf{Z}[\frac{1}{p}], j_* M_w)$ under the isomorphism (c).

The existence of $\zeta(M)$ having the relation with $\lim_{s\to 0} s^{-e} L(M,s)$ was conjectured by Beilinson [Be].

How zeta functions and p-adic zeta elements are related is illustrated in the following diagram.

$$\begin{array}{ccccc} \text{zeta functions side} & \text{(Betti)} & \xleftarrow{\text{Hodge theory}} & \text{(de Rham)} & \\ & \text{regulator}\uparrow & & \uparrow p\text{-adic Hodge theory} & \\ & (K\text{-theory}) & \xrightarrow[\text{Chern class}]{} & \text{(etale)} & p\text{-adic zeta elements side.} \end{array}$$

We have the following picture.

$$\begin{array}{ccccc} \text{automorphic rep} & \xleftarrow{?} & \text{motives} & \longrightarrow & p\text{-adic Gal rep} \\ \downarrow & & \downarrow ? & & \downarrow ? \\ \text{zeta functions} & & \text{rational zeta elements} & & p\text{-adic zeta elememts} \end{array}$$

The left upper arrow with a question mark shows the conjecture that the map {motives} $\to$ {zeta functions} factor through automorphic representations, which is a subject of non-abelian class field theory (Langlands correspondences). As the other question marks indicate, we do not know how to construct zeta elements in general, at present.

2.2. *p*-adic zeta elements for 1 dimensional galois representations. Let Λ be a commutative pro-p ring, and assume we are given a continuous homomorphism

$$\rho : \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \to GL_n(\Lambda)$$

which is unramified outside a finite set S of prime numbers S containing p. Let $\mathcal{F} = \Lambda^{\oplus n}$ on which $\mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q})$ acts via ρ, regarded as a sheaf on $\mathrm{Spec}(\mathbf{Z}[\frac{1}{S}])$ for the etale topology. We consider how to construct the p-adic zeta element $\zeta(\mathbf{Z}[\frac{1}{S}], \mathcal{F}, \Lambda)$.

In the case $n = 1$, we can use the "universal objects" as follows. Such ρ comes from the canonical homomorphism

$$\rho_{\mathrm{univ}} : \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \to GL_1(\Lambda_{\mathrm{univ}}) \quad \text{where} \quad \Lambda_{\mathrm{univ}} = \mathbf{Z}_p[[\mathrm{Gal}(\mathbf{Q}(\zeta_{Np^\infty})/\mathbf{Q})]]$$

for some $N \geq 1$ whose set of prime divisors coincide with S and for some continuous ring homomorphism $\Lambda_{\mathrm{univ}} \to \Lambda$. We have $\mathcal{F} \cong \mathcal{F}_{\mathrm{univ}} \otimes_{\Lambda_{\mathrm{univ}}} \Lambda$. Hence $\zeta(\mathbf{Z}[\frac{1}{S}], \mathcal{F}, \Lambda)$ should be defined to be the image of $\zeta(\mathbf{Z}[\frac{1}{S}], \mathcal{F}_{\mathrm{univ}}, \Lambda_{\mathrm{univ}})$. As is explained in [Ka$_2$] Ch. I, 3.3, $\zeta(\mathbf{Z}[\frac{1}{S}], \mathcal{F}_{\mathrm{univ}}, \Lambda_{\mathrm{univ}})$ is the pair of the p-adic Riemann zeta function and a system of cyclotomic units. Iwasawa main conjecture is regarded as the statemnet that this pair is a Λ_{univ}-basis of $\Delta(\mathbf{Z}[\frac{1}{S}], \mathcal{F}_{\mathrm{univ}}, \Lambda_{\mathrm{univ}})$.

2.3. *p*-adic zeta elements for 2 dimensional Galois representations. Now consider the case $n = 2$. The works of Hida, Wiles, and other people suggest that the universal objects Λ_{univ} and $\mathcal{F}_{\mathrm{univ}}$ for 2 dimensional Galois representations in which the determinant of the action of the complex conjugation is -1, are given by

$$\Lambda_{\mathrm{univ}} = \varprojlim_n \ p\text{-adic Hecke algebras of weight 2 and of level } Np^n,$$

$$\mathcal{F}_{univ} = \varprojlim_n \ H^1 \text{ of modular curves of level } Np^n.$$

Beilinson [Be] discovered ratinal zeta elements in K_2 of modular curves, and the images of these elements in the etale cohomology under the Chern class maps become p-adic zeta elements, and the inverse limit of these p-adic zeta elements should be $\zeta(\mathbf{Z}[\frac{1}{S}], \mathcal{F}_{\mathrm{univ}}, \Lambda_{\mathrm{univ}})$ at least conjecturally. By using this plan, the author obtained p-adic zeta elements for motives associated to eigen cusp forms of weight ≥ 2, from Beilinson elements. Here it is not yet proved that these p-adic zeta elements are actually basis of Δ, but it can be proved that they have the desired relations with values $L(E, \chi, 1)$ and $L(f, \chi, r)$ $(1 \leq r \leq k-1)$ for elliptic curves over $\mathbf{Q}$ (which are modular by [Wi], [BCDT]) and for eigen cusp forms of weight $k \geq 2$, and for Dirichlet charcaters χ. Beilinson elements are related in the Archimedean world

to $\lim_{s\to 0} s^{-1}L(E,\chi,s)$ for elliptic curves E over $\mathbf{Q}$, but not related to $L(E,\chi,1)$. However since they become universal (at least conjecturally) in the inverse limit in the p-adic world, we can obtain from them p-adic zeta elements related to $L(E,\chi,1)$. Using these elements and applying the method of Euler systems [Ko], [Pe_2], [Ru_2], [Ka_3], we can obtain the following results ([Ka_4]).

Theorem. *Let E be an elliptic curve over $\mathbf{Q}$, let $N \geq 1$, and let $\chi : Gal(\mathbf{Q}(\zeta_N)/\mathbf{Q}) \cong (\mathbf{Z}/N\mathbf{Z})^\times \to \mathbf{C}$ be a homomorphism. If $L(E,\chi,1) \neq 0$, the χ-part of $E(\mathbf{Q}(\zeta_N))$ and the χ part of the Tate-shafarevich group of E over $\mathbf{Q}(\zeta_N)$ are finite.*

The p-adic L-function $L_p(E)$ of E is constructed from the values $L(E,\chi,1)$.

Theorem. *Let E be an elliptic curve over $\mathbf{Q}$ which is of good reduction at p.*

(1) $rank(E(\mathbf{Q}) \leq ord_{s=1}L_p(E)$.

(2) Assume E is ordinary at p. Let $\Lambda = \mathbf{Z}_p[[Gal(\mathbf{Q}(\zeta_{p^\infty}/\mathbf{Q})]]$. Then the p-primary Selmer group of E over $\mathbf{Q}(\zeta_{p^\infty})$ is Λ-cotorsion and its characteristic polynomial divides $p^n L_p(E)$ for some n.

This result was proved by Rubin in the case of elliptic curves with complex multiplication ([Ru_1]).

As described above, we can obtain p-adic zeta elements of motives associated to eigen cusp forms of weight ≥ 2. For such modular forms, we can prove the analogous statement as the above (2).

Mazur and Greenberg conjectured that the charcteristic polynomial of the above p-primary Selmer group and the p-adic L-function divide each other.

References

[Be] Beilinson, A., Higher regulators and values of L-functions, *J. Soviet Math.*, 30 (1985), 2036–2070.

[BK] Bloch, S. and Kato, K., Tamagawa numbers of motives and L-functions, in The Grothendieck Festschrift, 1, Progress in Math., 86, Burkhauser (1990), 333–400.

[BCDT] Breuil, C., Conrad, B, Diamond, F., Taylor, R., On the modularity of elliptic curves over $\mathbf{Q}$: wild 3-adic exercises, *J. Amer. Math. Soc.*, 14 (2001), 834–939.

[FP] Fontaine, J. -M., and Perrin-Riou, B., Autour des conjectures de Bloch et Kato, cohomologie Galoisienne et valeurs de fonctions L, Proc. Symp. Pure Math. 55, Amer. Math. Soc., (1994), 599–706.

[Ka_1] Kato, K., Iwasawa theory and p-adic Hodge theory, *Kodai Math. J.*, 16 (1993), 1–31.

[Ka_2] Kato, K., Lectures on the approach to Iwasawa theory for Hasse-Weil L-functions via B_{dR}. I, Arithmetic algebraic geometry (Trento, 1991), 50–163, Lecture Notes in Math., 1553, Springer, Berlin (1993).

[Ka_3] Kato, K., Euler systems, Iwasawa theory, and Selmer groups, *Kodai Math. J.*, 22 (1999), 313–372.

[Ka_4] Kato, K., p-adic Hodge theory and values of zeta functions of modular forms, preprint.

[KM] Knudsen, F., and Mumford, D., The projectivity of the moduli space of stable curves I, *Math. Scand.*, 39, 1 (1976), 19–55.

[Ko] Kolyvagin, V. A., Euler systems, in The Grothendieck Festchrift, 2, Birkhouser (1990), 435–483.

[Pe_1] Perrin-Riou, B., Fonction L p-adiques des représentations p-adiques, Astérisque 229 (1995).

[Pe_2] Perrin-Riou, B., Systemes d'Euler p-adiques et théorie d'Iwaswa, *Ann. Inst. Fourier*, 48 (1998), 1231–1307.

[Ru_1] Rubin, K., The "main conjecture" of Iwasawa theory for imaginary quadratic fields, *Inventiones math.*, 103 (1991), 25–68.

[Ru_2] Rubin, K., Euler systems, Hermann Weyl Lectures, *Annals of Math. Studies*, 147, Princeton Univ. Press (2000).

[Wi] Wiles, A., Modular elliptic curves and Fermat's last theorem, *Ann. of Math.*, 141 (1995), 443–551.

ICM 2002 · Vol. II · 173–183

Derivatives of Eisenstein Series and Arithmetic Geometry*

Stephen S. Kudla†

Abstract

We describe connections between the Fourier coefficients of derivatives of Eisenstein series and invariants from the arithmetic geometry of the Shimura varieties M associated to rational quadratic forms (V, Q) of signature $(n, 2)$. In the case $n = 1$, we define generating series $\hat{\phi}_1(\tau)$ for 1-cycles (resp. $\hat{\phi}_2(\tau)$ for 0-cycles) on the arithmetic surface $\mathcal{M}$ associated to a Shimura curve over $\mathbb{Q}$. These series are related to the second term in the Laurent expansion of an Eisenstein series of weight $\frac{3}{2}$ and genus 1 (resp. genus 2) at the Siegel–Weil point, and these relations can be seen as examples of an 'arithmetic' Siegel–Weil formula. Some partial results and conjectures for higher dimensional cases are also discussed.

2000 Mathematics Subject Classification: 14G40, 14G35, 11F30.
Keywords and Phrases: Heights, Derivatives of Eisenstein series, Modular forms.

1. Introduction

In this report, we will survey results about generating functions for arithmetic cycles on Shimura varieties defined by rational quadratic forms of signature $(n, 2)$. For small values of n, these Shimura varieties are of PEL type, i.e., can be identified with moduli spaces for abelian varieties equipped with polarization, endomorphisms, and level structure. By analogy with CM or Heegner points on modular curves, cycles are defined by imposing additional endomorphisms. Relations between the heights or arithmetic degrees of such cycles and the Fourier coefficients of *derivatives* of Siegel Eisenstein series are proved in [10] and in subsequent joint work with Rapoport, [14], [15], [16], and with Rapoport and Yang [17], [18]. These relations may be viewed as an arithmetic version of the classical Siegel-Weil formula, which identifies the Fourier coefficients of *values* of Siegel Eisenstein

*Partially supported by NSF grant DMS-9970506 and by a Max-Planck Research Prize from the Max-Planck Society and Alexander von Humboldt Stiftung.

† Mathematics Department, University of Maryland, College Park, MD 20742, USA. E-mail: ssk@math.umd.edu

series with representation numbers of quadratic forms. The most complete example is that of anisotropic ternary quadratic forms ($n = 1$), so that the cycles are curves and 0-cycles on the arithmetic surfaces associated to Shimura curves. Other surveys of the material discussed here can be found in [11] and [12].

2. Shimura curves

Let B be an indefinite quaternion algebra over $\mathbb{Q}$, and let $D(B)$ be the product of the primes p for which $B_p = B \otimes_{\mathbb{Q}} \mathbb{Q}_p$ is a division algebra. The rational vector space

$$V = \{\ x \in B \mid \mathrm{tr}(x) = 0\ \}$$

with quadratic form given by $Q(x) = -x^2 = \nu(x)$, where $\mathrm{tr}(x)$ (resp. $\nu(x)$) is the reduced trace (resp. norm) of x, has signature $(1,2)$. The action of $B^\times$ on V by conjugation gives an isomorphism $G = \mathrm{GSpin}(V) \simeq B^\times$. Let

$$D = \{\ w \in V(\mathbb{C}) \mid (w,w) = 0,\ (w,\bar{w}) < 0\ \}/\mathbb{C}^\times \simeq \mathbb{P}^1(\mathbb{C}) \setminus \mathbb{P}^1(\mathbb{R})$$

be the associated symmetric space. Let O_B be a maximal order in B and let $\Gamma = O_B^\times$ be its unit group. The quotient $M(\mathbb{C}) = \Gamma\backslash D$ is the set of complex points of the Shimura curve M (resp. modular curve, if $D(B) = 1$) determined by B. This space should be viewed as an orbifold $[\Gamma\backslash D]$. For a more careful discussion of this and of the stack aspect, which we handle loosely here, see [18]. The curve M has a canonical model over $\mathbb{Q}$. From now on, we assume that $D(B) > 1$, so that M is projective. Drinfeld's model $\mathcal{M}$ for M over $\mathrm{Spec}\,(\mathbb{Z})$ is obtained as the moduli stack for abelian schemes (A, ι) with an action $\iota : O_B \hookrightarrow \mathrm{End}(A)$ satisfying the 'special' condition, [3]. It is proper of relative dimension 1 over $\mathrm{Spec}\,(\mathbb{Z})$, with semi-stable reduction at all primes and is smooth at all primes p at which B splits, i.e., for $p \nmid D(B)$. We view $\mathcal{M}$ as an arithmetic surface in the sense of Arakelov theory and consider its arithmetic Chow groups with real coefficients $\widehat{CH}^r(\mathcal{M}) = \widehat{CH}^r_{\mathbb{R}}(\mathcal{M})$, as defined in [2]. Recall that these groups are generated by pairs $(\mathcal{Z}, g)$, where $\mathcal{Z}$ is an $\mathbb{R}$-linear combination of divisors on $\mathcal{M}$ and g is a Green function for $\mathcal{Z}$, with relations given by $\mathbb{R}$-linear combinations of elements $\widehat{\mathrm{div}}\ (f) = (\mathrm{div}(f), -\log|f|^2)$ where $f \in \mathbb{Q}(\mathcal{M})^\times$ is a nonzero rational function on $\mathcal{M}$. These real vector spaces come equipped with a geometric degree map $\deg_{\mathbb{Q}} : \widehat{CH}^1(\mathcal{M}) \to CH^1(\mathcal{M}_{\mathbb{Q}}) \stackrel{\deg}{\to} \mathbb{R}$, where $\mathcal{M}_{\mathbb{Q}}$ is the generic fiber of $\mathcal{M}$, an arithmetic degree map $\widehat{\deg}\ : \widehat{CH}^2(\mathcal{M}) \to \mathbb{R}$, and the Gillet-Soulé height pairing, [2],

$$\langle\ ,\ \rangle : \widehat{CH}^1(\mathcal{M}) \times \widehat{CH}^1(\mathcal{M}) \longrightarrow \mathbb{R}.$$

Let $\mathcal{A}$ be the universal abelian scheme over $\mathcal{M}$. Then the Hodge line bundle $\omega = \epsilon^*(\Omega^2_{\mathcal{A}/\mathcal{M}})$ determined by $\mathcal{A}$ has a natural metric, normalized as in [18], section 3, and defines an element $\hat{\omega} \in \widehat{\mathrm{Pic}}(\mathcal{M})$, the group of metrized line bundles on

$\mathcal{M}$. We also write $\hat{\omega}$ for the image of this class in $\widehat{CH}^1(\mathcal{M})$ under the natural map, which sends a metrized line bundle $\hat{\mathcal{L}} = (\mathcal{L}, ||\ ||) \in \widehat{\mathrm{Pic}}(\mathcal{M})$ to the class of $(\mathrm{div}(s), -\log ||s||^2)$, for any nonzero section s of $\mathcal{L}$.

Arithmetic cycles in $\mathcal{M}$ are defined by imposing additional endomorphisms of the following type.

Definition 1. ([10]) *The space of special endomorphisms $V(A,\iota)$ of an abelian scheme (A,ι), as above, is*

$$V(A,\iota) = \{\ x \in \mathrm{End}(A) \mid x \circ \iota(b) = \iota(b) \circ x,\ \forall\, b \in O_B,\ \text{and } \mathrm{tr}(x) = 0\ \},$$

with $\mathbb{Z}$-valued quadratic form given by $-x^2 = Q(x)\,\mathrm{id}_A$.

2.1. Divisors

To obtain divisors on $\mathcal{M}$, we impose a single special endomorphism. For a positive integer t, let $\mathcal{Z}(t)$ be the divisor on $\mathcal{M}$ determined by the moduli stack of triples (A,ι,x) where (A,ι) is as before and where $x \in V(A,\iota)$ is a special endomorphism with $Q(x) = t$. Note that, for example, the complex points $\mathcal{Z}(t)(\mathbb{C})$ of $\mathcal{Z}(t)$ correspond to abelian surfaces (A,ι) over $\mathbb{C}$ with an 'extra' action of the order $\mathbb{Z}[\sqrt{-t}]$ in the imaginary quadratic field $\mathbb{Q}(\sqrt{-t})$, i.e., to CM points on the Shimura curve $M(\mathbb{C})$. On the other hand, the cycles $\mathcal{Z}(t)$ can have vertical components in the fibers of bad reduction $\mathcal{M}_p$ for $p \mid D(B)$. More precisely, in joint work with M. Rapoport we show:

Proposition 1. ([15]) *For $p \mid D(B)$, $\mathcal{Z}(t)$ contains components of the fiber of bad reduction $\mathcal{M}_p$ if and only if $\mathrm{ord}_p(t) \geq 2$ and no prime $\ell \mid D(B)$, $\ell \neq p$, splits in $\boldsymbol{k}_t := \mathbb{Q}(\sqrt{-t})$.*

The precise structure of the vertical part of $\mathcal{Z}(t)$ is determined in [15] using the Drinfeld-Cherednik p-adic uniformization of $\mathcal{M}_p$. For example, for $p \mid D(B)$, the multiplicities of the vertical components in the fiber $\mathcal{M}_p$ of the cycle $\mathcal{Z}(p^{2r}t)$ grow with r, while the horizontal part of this cycle remains unchanged.

To obtain classes in $\widehat{CH}^1(\mathcal{M})$, we construct Green functions by the procedure introduced in [10]. Let $L = O_B \cap V$. For $t \in \mathbb{Z}_{>0}$ and $v \in \mathbb{R}_{>0}$, define a function $\Xi(t,v)$ on $M(\mathbb{C})$ by

$$\Xi(t,v)(z) = \sum_{x \in L(t)} \beta_1(2\pi v R(x,z)),$$

where $L(t) = \{x \in L \mid Q(x) = t\}$, and, for $z \in D$ with preimage $w \in V(\mathbb{C})$, $R(x,z) = |(x,w)|^2\,|(w,\bar{w})|^{-1}$. Here

$$\beta_1(r) = \int_1^\infty e^{-ru}\,u^{-1}\,du = -\mathrm{Ei}(-r)$$

is the exponential integral. Recall that this function has a log singularity as r goes to zero and decays exponentially as r goes to infinity. In fact, as shown in [10], section 11, for any $x \in V(\mathbb{R})$ with $Q(x) \neq 0$, the function

$$\xi(x,z) := \beta_1(2\pi R(x,z))$$

can be viewed as a Green function on D for the divisor $D_x := \{z \in D \mid (x,z) = 0\}$. A simple calculation, [10], shows that, for $t > 0$, $\Xi(t,v)$ is a Green function of logarithmic type for the cycle $\mathcal{Z}(t)$, while, for $t < 0$, $\Xi(t,v)$ is a smooth function on $M(\mathbb{C})$.

Definition 2. *(i) For $t \in \mathbb{Z}$ and $v > 0$, the class $\widehat{\mathcal{Z}}(t,v) \in \widehat{CH}^1(\mathcal{M})$ is defined by:*

$$\widehat{\mathcal{Z}}(t,v) = \begin{cases} (\mathcal{Z}(t), \Xi(t,v)) & \text{if } t > 0, \\ -\hat{\omega} + (0, \mathbf{c} - \log(v)) & \text{if } t = 0, \\ (0, \Xi(t,v)) & \text{if } t < 0. \end{cases}$$

Here $\hat{\omega}$ is the metrized Hodge line bundle, as above, and the real constant $\mathbf{c}$ is given by

$$\frac{1}{2}\deg_{\mathbb{Q}}(\hat{\omega})\cdot\mathbf{c} = \langle\hat{\omega},\hat{\omega}\rangle - \zeta_{D(B)}(-1)\left[2\frac{\zeta'(-1)}{\zeta(-1)} + 1 - \log(4\pi) - \gamma - \sum_{p|D(B)}\frac{p\log(p)}{p-1}\right],$$

where $\zeta_{D(B)}(s) = \zeta(s)\prod_{p|D(B)}(1-p^{-s})$ and γ is Euler's constant.
(ii) For $\tau = u + iv \in \mathfrak{H}$ and $q = e(\tau) = e^{2\pi i\tau}$, the 'arithmetic theta function' $\hat{\phi}_1(\tau)$ is given by the generating series

$$\hat{\phi}_1(\tau) := \sum_{t\in\mathbb{Z}} \widehat{\mathcal{Z}}(t,v)\, q^t.$$

It is conjectured in [18] that the constant $\mathbf{c}$ occurring in the definition of $\widehat{\mathcal{Z}}(0,v)$ is, in fact, zero. It may be possible to use recent work of Bruinier and Kühn, [4], on the heights of curves on Hilbert modular surfaces to show that that $\langle\hat{\omega},\hat{\omega}\rangle$ has the predicted value and hence verify this conjecture.

Some justification for the terminology 'arithmetic theta function' is given by the following result, which is closely related to earlier work of Zagier, [25], and recent results of Borcherds, [1], cf. also [20].

Theorem 1. *The arithmetic theta function $\hat{\phi}_1(\tau)$ is a (nonholomorphic) modular form of weight $\frac{3}{2}$, valued in $\widehat{CH}^1(\mathcal{M})$, for a subgroup $\Gamma' \subset \mathrm{SL}_2(\mathbb{Z})$.*

The proof of Theorem 1 depends on Borcherd's result [1] and on the modularity of various complex valued q-expansions obtained by taking height pairings of $\hat{\phi}_1(\tau)$

with other classes in $\widehat{CH}^1(\mathcal{M})$. We now describe some of these in terms of values and derivatives of a certain Eisenstein series, [18], of weight $\frac{3}{2}$

$$\mathcal{E}_1(\tau, s, D(B)) = \sum_{\gamma \in \Gamma_\infty \backslash \mathrm{SL}_2(\mathbb{Z})} (c\tau + d)^{-\frac{3}{2}} |c\tau + d|^{-(s-\frac{1}{2})} v^{\frac{1}{2}(s-\frac{1}{2})} \Phi_1(s, \gamma, D(B)),$$

associated to B and the lattice L, and normalized so that it is invariant under $s \mapsto -s$. The main result of joint work with M. Rapoport and T. Yang is the following:

Theorem 2. ([18]) (i)

$$\mathcal{E}_1(\tau, \frac{1}{2}; D(B)) = \deg(\hat{\phi}_1(\tau)) = \sum_t \deg_{\mathbb{Q}}(\widehat{\mathcal{Z}}(t, v))\, q^t.$$

(ii)

$$\mathcal{E}_1'(\tau, \frac{1}{2}; D(B)) = \langle\, \hat{\phi}_1(\tau), \hat{\omega}\, \rangle = \sum_t \langle\, \widehat{\mathcal{Z}}(t, v), \hat{\omega}\, \rangle\, q^t.$$

Note that this result expresses the Fourier coefficients of the first two terms in the Laurent expansion at the point $s = \frac{1}{2}$ of the Eisenstein series $\mathcal{E}_1(\tau, s; D(B))$ in terms of the geometry and the arithmetic geometry of cycles on $\mathcal{M}$.

Next consider the image of

$$\hat{\phi}_1(\tau) - \mathcal{E}_1(\tau, \frac{1}{2}; D(B)) \cdot \deg(\hat{\omega})^{-1} \cdot \hat{\omega}$$

in $CH^1(\mathcal{M}_{\mathbb{Q}})$, the usual Chow group of the generic fiber. By (i) of Theorem 2, it lies in the Mordell-Weil space $CH^1(\mathcal{M}_{\mathbb{Q}})^0 \otimes \mathbb{C} \simeq \mathrm{Jac}(M)(\mathbb{Q}) \otimes_{\mathbb{Z}} \mathbb{C}$. In fact, it is essentially the generating function defined by Borcherds, [1], for the Shimura curve M, and hence is a holomorphic modular of weight $\frac{3}{2}$. For the case of modular curves, such a modular generating function, whose coefficients are Heegner points, was introduced by Zagier, [25]. By the Hodge index theorem for $\widehat{CH}^1(\mathcal{M})$, [2], the proof of Theorem 1 is completed by showing that the pairing of $\hat{\phi}_1(\tau)$ with each class of the form $(Y_p, 0)$, for Y_p a component of the fiber $\mathcal{M}_p$, $p \mid D(B)$ and each class of the form $(0, \phi)$, where $\phi \in C^\infty(M(\mathbb{C}))$, is modular.

2.2. 0-cycles

We next consider a generating function for 0-cycles on $\mathcal{M}$. Recall that the arithmetic Chow group $\widehat{CH}^2(\mathcal{M})$, with real coefficients, is generated by pairs $(\mathcal{Z}, g)$, where $\mathcal{Z}$ is a real linear combination of 0-cycles on $\mathcal{M}$ and g is a real smooth $(1,1)$-form on $\mathcal{M}(\mathbb{C})$. In fact, the arithmetic degree map, as defined in [2],

$$\widehat{\deg}\ : \widehat{CH}^2(\mathcal{M}) \to \mathbb{R}, \qquad \widehat{\deg}\ ((\mathcal{Z}, g)) = \sum_i n_i \log |k(P_i)| + \frac{1}{2} \int_{\mathcal{M}(\mathbb{C})} g,$$

where $\mathcal{Z} = \sum_i n_i P_i$ for closed points P_i of $\mathcal{M}$ with residue field $k(P_i)$, is an isomorphism.

Let $\tau = u + iv \in \mathfrak{H}_2$, the Siegel space of genus 2, and for $T \in \mathrm{Sym}_2(\mathbb{Z})$, let $q^T = e^{2\pi i \mathrm{tr}(T\tau)}$. To define the generating series

$$\hat{\phi}_2(\tau) = \sum_{T \in \mathrm{Sym}_2(\mathbb{Z})} \hat{\mathcal{Z}}(T, v)\, q^T,$$

we want to define classes $\hat{\mathcal{Z}}(T, v) \in \widehat{CH}^2(\mathcal{M})$ for each $T \in \mathrm{Sym}_2(\mathbb{Z})$ and $v \in \mathrm{Sym}_2(\mathbb{R})_{>0}$.

We begin by considering cycles on $\mathcal{M}$ which are defined by imposing pairs of endomorphisms. For $T \in \mathrm{Sym}_2(\mathbb{Z})_{>0}$ a positive definite integral symmetric matrix, let $\mathcal{Z}(T)$ be the moduli stack over $\mathcal{M}$ consisting of triples $(A, \iota, \mathbf{x})$ where (A, ι) is as before, and $\mathbf{x} = [x_1, x_2] \in V(A, \iota)^2$ is a pair of special endomorphisms with matrix of inner products $Q(\mathbf{x}) = \frac{1}{2}((x_i, x_j)) = T$. We call T the fundamental matrix of the triple $(A, \iota, \mathbf{x})$. The following result of joint work with M. Rapoport describes the cases in which $\mathcal{Z}(T)$ is, in fact, a 0-cycle on $\mathcal{M}$.

Proposition 2. ([15]) *Suppose that $T \in \mathrm{Sym}_2(\mathbb{Z})_{>0}$. (i) The cycle $\mathcal{Z}(T)$ is either empty or is supported in the set of supersingular points in a fiber $\mathcal{M}_p$ for a unique prime p determined by T. In particular, $\mathcal{Z}(T)_{\mathbb{Q}} = \emptyset$. The prime p is determined by the condition that T is represented by the ternary quadratic space $V^{(p)} = \{\, x \in B^{(p)} \mid \mathrm{tr}(x) = 0 \,\}$, with $Q^{(p)}(x) = -x^2$, where $B^{(p)}$ is the definite quaternion algebra over $\mathbb{Q}$ with $B_\ell^{(p)} \simeq B_\ell$ for all primes $\ell \neq p$. If there is no such prime, then $\mathcal{Z}(T)$ is empty.*
(ii) (T regular) *Let p be as in (i). Then, if $p \nmid D(B)$ or if $p \mid D(B)$ but $p^2 \nmid T$, then $\mathcal{Z}(T)$ is a 0-cycle in $\mathcal{M}_p$.*
(iii) (T irregular) *Let p be as in (i). If $p \mid D(B)$ and $p^2 \mid T$, then $\mathcal{Z}(T)$ is a union, with multiplicities, of components of $\mathcal{M}_p$, cf.* [15], *176.*

For $T \in \mathrm{Sym}_2(\mathbb{Z})_{>0}$ regular, as in (ii) of Proposition 2, we let

$$\hat{\mathcal{Z}}(T, v) := \hat{\mathcal{Z}}(T) = (\mathcal{Z}(T), 0) \in \widehat{CH}^2(\mathcal{M}).$$

For $T = \begin{pmatrix} t_1 & m \\ m & t_2 \end{pmatrix} \in \mathrm{Sym}_2(\mathbb{Z})_{>0}$ irregular, we use the results of [15], section 8 (where the quadratic form is taken with the opposite sign). We must therefore assume that $p \neq 2$, although the results of the appendix to section 11 of [18] suggest that it should be possible to eliminate this restriction. In this case, the vertical cycle $\mathcal{Z}(T)$ in the fiber $\mathcal{M}_p$ is the union of those connected components of the intersection $\mathcal{Z}(t_1) \times_{\mathcal{M}} \mathcal{Z}(t_2)$ where the 'fundamental matrix', [15], is equal to T. Here $\mathcal{Z}(t_1)$ and $\mathcal{Z}(t_2)$ are the codimension 1 cycles defined earlier. Note that, by Proposition 1, they can share some vertical components. We base change to $\mathbb{Z}_p$ and set

$$\hat{\mathcal{Z}}(T, v) := \chi(\mathcal{Z}(T), \mathcal{O}_{\mathcal{Z}(t_1)} \otimes^{\mathbb{L}} \mathcal{O}_{\mathcal{Z}(t_2)}) \cdot \log(p) \quad \in \mathbb{R} \simeq \widehat{CH}^2(\mathcal{M}),$$

where χ is the Euler-Poincaré characteristic of the derived tensor product of the structure sheaves $\mathcal{O}_{\mathcal{Z}(t_1)}$ and $\mathcal{O}_{\mathcal{Z}(t_2)}$, cf. [15], section 4. Note that the same definition could have been used in the regular case.

Next we consider nonsingular $T \in \mathrm{Sym}_2(\mathbb{Z})$ of signature $(1,1)$ or $(0,2)$. In this case, $\mathcal{Z}(T)$ is empty, since the quadratic form on $V(A,\iota)$ is positive definite, and our 'cycle' should be viewed as 'vertical at infinity'. For a pair of vectors $\mathbf{x} = [x_1, x_2] \in V(\mathbb{Q})^2$ with nonsingular matrix of inner products $Q(\mathbf{x}) = \frac{1}{2}\big((x_i, x_j)\big)$, the quantity

$$\Lambda(\mathbf{x}) := \int_D \xi(x_1) * \xi(x_2),$$

where $\xi(x_1) * \xi(x_2)$ is the $*$-product of the Green functions $\xi(x_1)$ and $\xi(x_2)$, [6], is well defined and depends only on $Q(\mathbf{x})$. In addition, $\Lambda(\mathbf{x})$ has the following remarkable invariance property.

Theorem 3. ([10, Theorem 11.6]) *For $k \in O(2)$, $\Lambda(\mathbf{x} \cdot k) = \Lambda(x)$.*

For $T \in \mathrm{Sym}_2(\mathbb{Z})$ of signature $(1,1)$ or $(0,2)$ and for $v \in \mathrm{Sym}_2(\mathbb{R})_{>0}$, choose $a \in GL_2(\mathbb{R})$ such that $v = a\,{}^t a$, and define

$$\widehat{\mathcal{Z}}(T,v) := \sum_{\mathbf{x} \in L^2,\ Q(\mathbf{x}) = T,\ \bmod \Gamma} \Lambda(\mathbf{x}\,a) \quad \in \mathbb{R} \simeq \widehat{CH}^2(\mathcal{M}).$$

Here $L = O_B \cap V$ and $\Gamma = O_B^\times$, as before. Note that the invariance property of Theorem 3 is required to make the right side independent of the choice of a.

We omit the definition of the terms for singular T's, cf. [11].

By analogy with Theorem 1, we conjecture that, with this definition, the generating series $\hat{\phi}_2(\tau)$ is the q-expansion of a Siegel modular form of weight $\frac{3}{2}$ for a subgroup $\Gamma' \subset \mathrm{Sp}_2(\mathbb{Z})$. More precisely, there is a normalized Siegel Eisenstein series $\mathcal{E}_2(\tau, s; D(B))$ of weight $\frac{3}{2}$ attached to B, [10].

Conjecture 1.

$$\mathcal{E}_2'(\tau, 0; D(B)) \stackrel{?}{=} \hat{\phi}_2(\tau). \tag{$C1$}$$

This amounts to the family of identities

$$\mathcal{E}_{2,T}'(\tau, 0; D(B)) \stackrel{?}{=} \widehat{\mathcal{Z}}(T,v)\, q^T \tag{$C1_T$}$$

on Fourier coefficients, for all $T \in \mathrm{Sym}_2(\mathbb{Z})$. Here the isomorphism $\widehat{\deg}$ is being used.

Theorem 4. ([10], [15]) *The Fourier coefficient identity $(C1_T)$ holds in the following cases:*
(i) $T \in \mathrm{Sym}_2(\mathbb{Z})$ is not represented by V or by any of the spaces $V^{(p)}$ of Proposition 2.
(In this case both $\widehat{\mathcal{Z}}(T,v)$ and $\mathcal{E}_{2,T}'(\tau, 0; D(B))$ are zero.)

(*ii*) $T \in \mathrm{Sym}_2(\mathbb{Z})_{>0}$ *is regular and* $p \nmid 2D(B)$, [10].
(*iii*) $T \in \mathrm{Sym}_2(\mathbb{Z})_{>0}$ *is irregular with* $p \neq 2$, *or regular with* $p \mid D(B)$ *and* $p \neq 2$, [15].
(*iv*) $T \in \mathrm{Sym}_2(\mathbb{Z})$ is nonsingular of signature $(1,1)$ or $(0,2)$, [10].

Theorem 4 is proved by a direct computation of both sides of $(C1_T)$. In case (ii), the computation of the Fourier coefficient $\mathcal{E}'_{2,T}(\tau, 0; D(B))$ depends on the formula of Kitaoka, [8], for the local representation densities $\alpha_p(S,T)$ for the given T and a variable unimodular S. The computation of $\widehat{\mathcal{Z}}(T,v) = \widehat{\deg}\,((\mathcal{Z}(T),0))$ depends on a special case of a result of Gross and Keating, [7], about the deformations of a triple of isogenies between a pair of p-divisible formal groups of dimension 1 and height 2 over $\bar{\mathbb{F}}_p$. Their result is also valid for $p = 2$, so it should be possible to extend (ii) to the case $p = 2$ by extending the result of Kitaoka.

In case (iii), an explicit formula for the quantity $\chi(\mathcal{Z}(T), \mathcal{O}_{\mathcal{Z}(t_1)} \otimes^{\mathbb{L}} \mathcal{O}_{\mathcal{Z}(t_2)})$ is obtained in [15] using p-adic uniformization. The analogue of Kitaoka's result is a determination of $\alpha_p(S,T)$ for arbitrary S due to T. Yang, [22]. In both of these results, the case $p = 2$ remains to be done.

Case (iv) is proved by directly relating the function Λ, defined via the $*$-product to the derivative at $s = 0$ of the confluent hypergeometric function of a matrix argument defined by Shimura, [21]. The invariance property of Theorem 3 plays an essential role. The case of signature $(1,1)$ is done in [10]; the argument for signature $(0,2)$ is the same.

A more detailed sketch of the proofs can be found in [11].

As part of ongoing joint work with M. Rapoport and T. Yang, the verification of $(C1_T)$ for singular T of rank 1 is nearly complete.

3. Higher dimensional examples

So far, we have discussed the generating functions $\hat{\phi}_1(\tau) \in \widehat{CH}^1(\mathcal{M})$ and $\hat{\phi}_2(\tau) \in \widehat{CH}^2(\mathcal{M})$ attached to the arithmetic surface $\mathcal{M}$, and the connections of these series to derivatives of Eisenstein series. There should be analogous series defined as generating functions for arithmetic cycles for the Shimura varieties attached to rational quadratic spaces (V,Q) of signature $(n,2)$. At present there are several additional examples, all based on the accidental isomorphisms for small values of n, which allow us to identify the Shimura varieties in question with moduli spaces of abelian varieties with specified polarization and endomorphisms. Here we briefly sketch what one hopes to obtain and indicate what is known so far. The results here are joint work with M. Rapoport.

Hilbert-Blumenthal varieties $(n = 2)$**,** [14]**.** When the rational quadratic space (V,Q) has signature $(2,2)$, the associated Shimura variety M is a quasi-projective surface with a canonical model over $\mathbb{Q}$. There is a model $\mathcal{M}$ of M over $\mathrm{Spec}\,(\mathbb{Z}[N^{-1}])$ defined as the moduli scheme for collections $(A, \lambda, \iota, \bar{\eta})$ where A is an abelian scheme of relative dimension 8 dimension with polarization λ, level

structure $\bar{\eta}$, and an action of $O_C \otimes \mathcal{O}_{\boldsymbol{k}}$, where O_C is a maximal order in the Clifford algebra $C(V)$ of V and $\mathcal{O}_{\boldsymbol{k}}$ is the ring of integers in the quadratic field $\boldsymbol{k} = \mathbb{Q}(\sqrt{d})$ for $d = \mathrm{discr(V)}$, the discriminant field of V, [14]. Again, a space $V(A,\iota) = V(A,\lambda,\iota,\bar{\eta})$ of special endomorphisms is defined; it is a $\mathbb{Z}$-module of finite rank equipped with a positive definite quadratic form Q. For $T \in \mathrm{Sym}_r(\mathbb{Z})$, we let $\mathcal{Z}(T)$ be the locus of $(A,\lambda,\iota,\bar{\eta},\mathbf{x})$'s where $\mathbf{x} = [x_1,\ldots,x_r]$, $x_i \in V(A,\iota)$ is a collection of r special endomorphisms with matrix of inner products $Q(\mathbf{x}) = \frac{1}{2}\big((x_i,x_j)\big) = T$.

One would *like* to define a family of generating functions according to the following conjectural chart. Again there is a metrized Hodge line bundle $\hat{\omega} \in \widehat{CH}^1(\mathcal{M})$.

$r=1$, $\mathcal{Z}(t)_{\mathbb{Q}} = \mathrm{HZ-curve}$, $\hat{\phi}_1(\tau) = \hat{\omega} + ? + \sum_{t\neq 0} \widehat{\mathcal{Z}}(t,v)\, q^t$, $\langle \hat{\phi}_1(\tau), \hat{\omega}^2\rangle \stackrel{?}{=} \mathcal{E}_1'(\tau,1)$.

$r=2$, $\mathcal{Z}(t)_{\mathbb{Q}} = 0\text{-cycle}$, $\hat{\phi}_2(\tau) = \hat{\omega}^2 + ? + \sum_{T\neq 0} \widehat{\mathcal{Z}}(T,v)\, q^T \langle \hat{\phi}_2(\tau), \hat{\omega}\rangle \stackrel{?}{=} \mathcal{E}_2'(\tau,\frac{1}{2})$.

$r=3$, $\mathcal{Z}(T)_{\mathbb{Q}} = \emptyset$, $\hat{\phi}_3(\tau) = \hat{\omega}^3 + ? + \sum_{T\neq 0} \widehat{\mathcal{Z}}(T,v)\, q^T \widehat{\deg}\, \hat{\phi}_3(\tau) \stackrel{?}{=} \mathcal{E}_3'(\tau,0)$.

Here, the generating function $\hat{\phi}_r(\tau)$ is valued in $\widehat{CH}^r(\mathcal{M})$, the rth arithmetic Chow group, $\mathcal{E}_r(\tau,s)$ is a certain normalized Siegel Eisenstein series of genus r, and the critical value of s in the identity in the last column is the Siegel-Weil point $s_0 = \frac{1}{2}(\dim(V) - r - 1)$. Of course, one would like the $\hat{\phi}_r(\tau)$'s to be Siegel modular forms of genus r and weight 2.

There are many technical problems which must be overcome to obtain such results. For example, one would like to work with a model over $\mathrm{Spec}\,(\mathbb{Z})$. If V is anisotropic, then M is projective, but if V is isotropic, e.g., for the classical Hilbert-Blumenthal surfaces where it has $\mathbb{Q}$-rank 1, then one must compactify. Since the metric on $\hat{\omega}$ is singular at the boundary a more general version of the Gillet-Soulé theory, currently being developed by Burgos, Kramer and Kühn, [5], [19], will be needed.

Nonetheless, the chart suggests many identities which can in fact be checked rigorously. For example, there are again rational quadratic spaces $V^{(p)}$ of dimension 4 and signature $(4,0)$ obtained by switching the Hasse invariant of V at p.

Theorem 5. [14], [11]. *(i) If $T \in \mathrm{Sym}_3(\mathbb{Z})_{>0}$ is not represented by any of the $V^{(p)}$'s, then $\mathcal{Z}(T) = \emptyset$ and $\mathcal{E}'_{3,T}(\tau,0) = 0$.*
(ii) If $T \in \mathrm{Sym}_3(\mathbb{Z})_{>0}$ is represented by $V^{(p)}$ where p is a prime of good reduction split in $\boldsymbol{k}$, then $\mathcal{Z}(T)$ is a 0-cycle in $\mathcal{M}_p$ and

$$\widehat{\deg}\,((\mathcal{Z}(T),0))\, q^T = \mathcal{E}'_{3,T}(\tau,0). \qquad (\star)$$

(iii) If $T \in \mathrm{Sym}_3(\mathbb{Z})_{>0}$ is represented by $V^{(p)}$ and p is a prime of good reduction inert in $\boldsymbol{k}$, then $\mathcal{Z}(T)$ is a 0-cycle in $\mathcal{M}_p$ if and only if $p \nmid T$. If this is the case, then the Fourier coefficient identity $(\star)$ again holds. If $p \mid T$, then $\mathcal{Z}(T)$ is a union of components of the supersingular locus of $\mathcal{M}_p$.

Finally, say if V is anisotropic, one can consider the image $\mathrm{cl}(\hat{\phi}_r(\tau)) \in H^{2r}(M,\mathbb{C})$ of $\hat{\phi}_r(\tau)$ in the usual (Betti) cohomology of $\mathcal{M}(\mathbb{C})$. Of course, $\mathrm{cl}(\hat{\phi}_3(\tau)) = 0$ for degree reasons. Joint work with J. Millson on generating functions for cohomology classes of special cycles yields:

Theorem 6. ([13], [9], [11]) *Suppose that V is anisotropic.* (*i*) $\mathrm{cl}(\hat\phi_r(\tau))$ *is a Siegel modular form of genus r and weight 2 valued in $H^{2r}(M,\mathbb{C})$.*
(*ii*) *For the cup product pairing,* $\big(\,\mathrm{cl}(\hat\phi_r(\tau)), \mathrm{cl}(\hat\omega)\,\big) = \mathcal{E}_r(\tau, s_0)$, *where* $s_0 = \frac{1}{2}(3-r)$.

Part (ii) here generalizes (i) of Theorem 2 above, so that, again, the value at s_0 of the Eisenstein series $\mathcal{E}_r(\tau,s)$ involves the complex geometry, while, conjecturally, the second term involves the height pairing.

Siegel modular varieties ($n = 3$), [16]. Here, an integral model $\mathcal{M}$ of the Shimura variety M attached to a rational quadratic space of signature $(3,2)$ can be obtained as a moduli space of polarized abelian varieties of dimension 16 with an action of a maximal order O_C in the Clifford algebra of V. We just give the relevant *conjectural* chart:

$$r=1,\quad \mathcal{Z}(t)_{\mathbb{Q}} = \begin{matrix}\text{Humbert}\\ \text{surface}\end{matrix},\quad \hat\phi_1(\tau) = \hat\omega + ? + \textstyle\sum_{t\neq 0} \widehat{\mathcal{Z}}(t,v)\, q^t,\quad \langle \hat\phi_1(\tau), \hat\omega^3\rangle \stackrel{?}{=} \mathcal{E}_1'(\tau, \tfrac{3}{2}).$$
$$r=2,\quad \mathcal{Z}(t)_{\mathbb{Q}} = \text{curve}\qquad\quad \hat\phi_2(\tau) = \hat\omega^2 + ? + \textstyle\sum_{T\neq 0} \widehat{\mathcal{Z}}(T,v)\, q^T \langle \hat\phi_2(\tau), \hat\omega^2\rangle \stackrel{?}{=} \mathcal{E}_2'(\tau, 1).$$
$$r=3,\quad \mathcal{Z}(T)_{\mathbb{Q}} = \text{0-cycle},\quad \hat\phi_3(\tau) = \hat\omega^3 + ? + \textstyle\sum_{T\neq 0} \widehat{\mathcal{Z}}(T,v)\, q^T \langle \hat\phi_2(\tau), \hat\omega\rangle \stackrel{?}{=} \mathcal{E}_3'(\tau, \tfrac{1}{2}).$$
$$r=4,\quad \mathcal{Z}(T)_{\mathbb{Q}} = \emptyset,\qquad\quad \hat\phi_4(\tau) = \hat\omega^4 + ? + \textstyle\sum_{T\neq 0} \widehat{\mathcal{Z}}(T,v)\, q^T \widehat{\deg}\, \hat\phi_4(\tau) \stackrel{?}{=} \mathcal{E}_4'(\tau, 0).$$

Here the Eisenstein series and, conjecturally, the generating functions $\hat\phi_r(\tau)$ have weight $\frac{5}{2}$, and the values of the Eisenstein series should be related to the series $\mathrm{cl}(\hat\phi_r(\tau))$. In the case of a prime p of good reduction a model of M over $\mathrm{Spec}\,(\mathbb{Z}_p)$ is defined in [16], and cycles are defined by imposing special endomorphisms. For example, for $r=4$, the main results of [16] give a criterion for $\mathcal{Z}(T)$ to be a 0-cycle in a fiber $\mathcal{M}_p$ and show that, when this is the case, then $\widehat{\deg}((\mathcal{Z}(T),0))\, q^T = \mathcal{E}_{4,T}'(\tau,0)$. The calculation of the left hand side is again based on the result of Gross and Keating mentioned in the description of the proof of Theorem 4 above. This provides some evidence for the last of the derivative identities in the chart.

References

[1] R. Borcherds, The Gross-Kohnen-Zagier theorem in higher dimensions, *Duke Math. J.* **97** (1999), 219–233.

[2] J.-B. Bost, Potential theory and Lefschetz theorems for arithmetic surfaces, *Ann. Sci. École Norm. Sup.* **32** (1999), 241–312.

[3] J.-F. Boutot and H. Carayol, Uniformisation p-adique des courbes de Shimura, in Astérisque, vol. **196–197**, 1991, pp. 45–158.

[4] J. H. Bruinier, J. Burgos, and U. Kühn, in preparation.

[5] J. Burgos, J. Kramer and U. Kühn, in preparation.

[6] H. Gillet and C. Soulé, Arithmetic intersection theory, *Publ. Math. IHES* **72** (1990), 93–174.

[7] B. H. Gross and K. Keating, On the intersection of modular correspondences, *Invent. Math.* **112** (1993), 225–245.

[8] Y. Kitaoka, A note on local densities of quadratic forms, *Nagoya Math. J.* **92** (1983), 145–152.

[9] S. Kudla, Algebraic cycles on Shimura varieties of orthogonal type, *Duke Math. J.* **86** (1997), 39–78.

[10] ———, Central derivatives of Eisenstein series and height pairings, *Ann. of Math.* **146** (1997), 545-646.

[11] ———, Derivatives of Eisenstein series and generating functions for arithmetic cycles, Sém. Bourbaki n^o 876, Astérisque, vol. **276**, 2002, pp. 341–368.

[12] ———, Special cycles and derivatives of Eisenstein series, *Proc. of MSRI Workshop on Heegner points (to appear).*

[13] S. Kudla and J. Millson, Intersection numbers of cycles on locally symmetric spaces and Fourier coefficients of holomorphic modular forms in several complex variables, *Publ. Math. IHES* **71** (1990), 121–172.

[14] S. Kudla and M. Rapoport, Arithmetic Hirzebruch–Zagier cycles, *J. reine angew. Math.* **515** (1999), 155–244.

[15] ———, Height pairings on Shimura curves and p-adic unformization, *Invent. math.* **142** (2000), 153–223.

[16] ———, Cycles on Siegel threefolds and derivatives of Eisenstein series, *Ann. Scient. Éc. Norm. Sup.* **33** (2000), 695–756.

[17] S. Kudla, M. Rapoport and T. Yang, On the derivative of an Eisenstein series of weight 1, *Int. Math. Res. Notices, No.7* (1999), 347–385.

[18] ———, Derivatives of Eisenstein series and Faltings heights, *preprint* (2001).

[19] U. Kühn, Generalized arithmetic intersection numbers, *J. reine angew. Math.* **534** (2001), 209–236.

[20] W. J. McGraw, On the rationality of vector-valued modular forms, *preprint* (2001).

[21] G. Shimura, Confluent hypergeometric functions on tube domains, *Math. Annalen* **260** (1982), 269–302.

[22] T. Yang, An explicit formula for local densities of quadratic forms, *J. Number Theory* **72** (1998), 309–356.

[23] ———, The second term of an Eisenstein series, *Proc. of the ICCM, (to appear).*

[24] ———, Faltings heights and the derivative of Zagier's Eisenstein series, *Proc. of MSRI workshop on Heegner points, preprint (2002).*

[25] D. Zagier, Modular points, modular curves, modular surfaces and modular forms, Lecture Notes in Math. 1111, Springer, Berlin, 1985, 225–248.

ICM 2002 · Vol. II · 185–195

Elliptic Curves and Class Field Theory

Barry Mazur* **Karl Rubin**†

Abstract

Suppose E is an elliptic curve defined over $\mathbf{Q}$. At the 1983 ICM the first author formulated some conjectures that propose a close relationship between the explicit class field theory construction of certain abelian extensions of imaginary quadratic fields and an explicit construction that (conjecturally) produces almost all of the rational points on E over those fields.

Those conjectures are to a large extent settled by recent work of Vatsal and of Cornut, building on work of Kolyvagin and others. In this paper we describe a collection of interrelated conjectures still open regarding the variation of Mordell-Weil groups of E over abelian extensions of imaginary quadratic fields, and suggest a possible algebraic framework to organize them.

2000 Mathematics Subject Classification: 11G05, 11R23.
Keywords and Phrases: Elliptic curves, Iwasawa theory, Heegner points.

1. Introduction

Eighty years have passed since Mordell proved that the (Mordell-Weil) group of rational points on an elliptic curve E is finitely generated, yet so limited is our knowledge that we still have no algorithm guaranteed to compute the rank of this group. If we want to ask even more ambitious questions about how the rank of the Mordell-Weil group $E(F)$ varies as F varies, it makes sense to restrict attention only to those fields for which we have an explicit construction, such as finite abelian extensions of a given imaginary quadratic field K. Taking our lead from the profound discovery of Iwasawa that the variational properties of certain arithmetic invariants are well-behaved if one restricts to subfields of $\mathbf{Z}_p^d$-extensions of number fields, we will focus on the following Mordell-Weil variation problem:

Fixing an elliptic curve E defined over $\mathbf{Q}$, an imaginary quadratic field K, and a prime number p, study the variation of the Mordell-Weil group of E over finite subfields of the (unique) $\mathbf{Z}_p^2$-extension of K in $\bar{K}$.

*Department of Mathematics, Harvard University, Cambridge, MA 02138, USA. E-mail: mazur@math.harvard.edu
†Department of Mathematics, Stanford University, Stanford, CA 94305, USA. E-mail: rubin@math.stanford.edu

This problem was the subject of some conjectures formulated by the first author at the 1983 ICM [8], conjectures which have recently been largely settled by work of Vatsal [15] and Cornut [1] building on work of Kolyvagin and others.

Example. Let E be the elliptic curve $y^2+y=x^3-x$, $p=5$, and let $K=\mathbf{Q}(\sqrt{-7})$. If F is a finite extension of K, contained in the $\mathbf{Z}_5^2$ extension of K, then $\operatorname{rank} E(F) = [F\cap K_\infty^{\mathrm{anti}} : K]$ where K_∞^{anti} is the anticyclotomic $\mathbf{Z}_5$-extension of K (see §2 for the definition). One only has an answer like this in the very simplest cases.

Now with the same E and p, take $K=\mathbf{Q}(\sqrt{-26})$. A guess here would be that $\operatorname{rank} E(F) = [F\cap K_\infty^{\mathrm{anti}} : K]+2$, but this seems to be beyond present technology.

The object of this article is to sketch a package of still-outstanding conjectures in hopes that it offers an even more precise picture of this piece of arithmetic. These conjectures are in some cases due to, and in other cases build on ideas of, Bertolini & Darmon, Greenberg, Gross & Zagier, Haran, Hida, Iwasawa, Kolyvagin, Nekovář, Perrin-Riou, and the authors, among others.

In sections 3 through 5 we describe the three parts of our picture: the *arithmetic theory* (the study of the Selmer modules over Iwasawa rings that contain the information we seek), the *analytic theory* (the construction and study of the relevant L-functions, both classical and p-adic), and the *universal norm theory* which arises from purely arithmetic considerations, but provides analytic invariants.

In the final section we suggest the beginnings of a new algebraic structure to organize these conjectures. This structure should not be viewed as a conjecture, but rather as a mnemonic to collect our conjectures and perhaps predict new ones.

More details and proofs will appear in a forthcoming paper.

2. Running hypotheses and notation

Fix a triple (E,K,p) where E is an elliptic curve of conductor N over $\mathbf{Q}$, K is an imaginary quadratic field of discriminant $D<-4$, and p is a prime number. To keep our discussion focused and as succinct as possible, we make the following hypotheses and conventions.

Assume that p is odd, that N, p and D are pairwise relatively prime, and that if E has complex multiplication, then K is *not* its field of complex multiplication. Let $\mathcal{O}_K\subset K$ denote the ring of integers of K. Assume further that there exists an ideal $\mathcal{N}\subset\mathcal{O}_K$ such that $\mathcal{O}_K/\mathcal{N}$ is cyclic of order N (this is sometimes called the **Heegner Hypothesis**), and that p is a prime of ordinary reduction for E. For simplicity we will assume throughout this article that the p-primary subgroups of the Shafarevich-Tate groups of E over the number fields we consider are all finite.

Proposition 1. *Under the assumptions above,* $\operatorname{rank} E(K)$ *is odd.*

Proof. This follows from the Parity Conjecture recently proved by Nekovář [11].

Let $\mathbf{K}_\infty$ denote the (unique) $\mathbf{Z}_p^2$-extension of K and $\mathbf{\Gamma} := \operatorname{Gal}(\mathbf{K}_\infty/K)$, so $\mathbf{\Gamma}\cong\mathbf{Z}_p^2$. We define the Iwasawa ring

$$\mathbf{\Lambda} := \mathbf{Z}_p[[\mathbf{\Gamma}]]\otimes_{\mathbf{Z}_p}\mathbf{Q}_p.$$

(To simplify notation and to avoid some complications, we will often work with $\mathbf{Q}_p$-vector spaces instead of natural $\mathbf{Z}_p$-modules; in particular we have tensored the usual Iwasawa ring with $\mathbf{Q}_p$.) For every (finite or infinite) extension F of K in $\mathbf{K}_\infty$ we also define

$$\Lambda_F := \mathbf{Z}_p[[\mathrm{Gal}(F/K)]] \otimes_{\mathbf{Z}_p} \mathbf{Q}_p, \qquad \mathbf{I}_F := \ker\{\mathbf{\Lambda} \twoheadrightarrow \Lambda_F\}.$$

Then $\mathbf{I}_K$ is the augmentation ideal of $\mathbf{\Lambda}$, and if $[F : K]$ is finite then Λ_F is just the group ring $\mathbf{Q}_p[\mathrm{Gal}(F/K)]$. If $\mathrm{Gal}(F/K)$ is $\mathbf{Z}_p$ or $\mathbf{Z}_p^2$, and M is a finitely generated torsion Λ_F-module, then $\mathrm{char}_{\Lambda_F}(M)$ will denote the characteristic ideal of M. In particular $\mathrm{char}_{\Lambda_F}(M)$ is a principal ideal of Λ_F.

There is a $\mathbf{Q}_p$-projective line of $\mathbf{Z}_p$-extensions of K, all contained in $\mathbf{K}_\infty$. Among these are two distinguished $\mathbf{Z}_p$-extensions:

- the **cyclotomic** $\mathbf{Z}_p$-extension K_∞^{cycl}, the compositum of K with the unique (cyclotomic) $\mathbf{Z}_p$-extension of $\mathbf{Q}$ (write $\Gamma_{\mathrm{cycl}} = \mathrm{Gal}(K_\infty^{\mathrm{cycl}}/K)$, $\Lambda_{\mathrm{cycl}} = \Lambda_{K_\infty^{\mathrm{cycl}}}$),
- the **anticyclotomic** $\mathbf{Z}_p$-extension K_∞^{anti}, the unique $\mathbf{Z}_p$-extension of K that is Galois over $\mathbf{Q}$ with non-abelian, and in fact dihedral, Galois group (write $\Gamma_{\mathrm{anti}} = \mathrm{Gal}(K_\infty^{\mathrm{anti}}/K)$, $\Lambda_{\mathrm{anti}} = \Lambda_{K_\infty^{\mathrm{anti}}}$).

Then $\mathbf{\Gamma} = \Gamma_{\mathrm{cycl}} \oplus \Gamma_{\mathrm{anti}}$ and $\mathbf{\Lambda} = \Lambda_{\mathrm{cycl}} \otimes_{\mathbf{Z}_p} \Lambda_{\mathrm{anti}}$.

Complex conjugation $\tau : K \to K$ acts on $\mathbf{\Gamma}$, acting as $+1$ on Γ_{cycl} and -1 on Γ_{anti}. This induces nontrivial involutions of $\mathbf{\Lambda}$ and Λ_{anti}, which we also denote by τ. If M is a module over $\mathbf{\Lambda}$ (or similarly over Λ_{anti}), let $M^{(\tau)}$ denote the module whose underlying abelian group is M but where the new action of $\gamma \in \mathbf{\Gamma}$ on $m \in M^{(\tau)}$ is given by the old action of γ^τ on m.

Our $\mathbf{\Lambda}$-modules will usually come with a natural action of $\mathrm{Gal}(\mathbf{K}_\infty/\mathbf{Q})$. These actions are continuous and $\mathbf{Z}_p$-linear, and satisfy the formula $\tilde{\tau}(\gamma \cdot m) = \gamma^\tau \cdot \tilde{\tau}(m)$ for every lift $\tilde{\tau}$ of τ to $\mathrm{Gal}(\mathbf{K}_\infty/\mathbf{Q})$. Thus the action of any lift $\tilde{\tau}$ induces an isomorphism $M \xrightarrow{\sim} M^{(\tau)}$. We will refer to such $\mathbf{\Lambda}$ or Λ_{anti}-modules as **semi-linear τ-modules**. If M is a semi-linear τ-module and is free of rank one over Λ_{anti}, we define the **sign** of M to be the sign ± 1 of the action of τ on the one-dimensional $\mathbf{Q}_p$-vector space $M \otimes_{\Lambda_{\mathrm{anti}}} \Lambda_K$. Such an M is completely determined (up to isomorphism preserving its structure) by its sign.

Definition 2. *If M and A are semi-linear τ-modules, then a ($\mathbf{\Lambda}$-bilinear) A-valued* **τ-Hermitian pairing** *π is a $\mathbf{\Lambda}$-module homomorphism $\pi : M \otimes_{\mathbf{\Lambda}} M^{(\tau)} \to A$ such that for every lift $\tilde{\tau}$ of τ to* $\mathrm{Gal}(\mathbf{K}_\infty/\mathbf{Q})$

$$\pi(m \otimes n) = \pi(n \otimes m)^{\tilde{\tau}} = \pi(\tilde{\tau}n \otimes \tilde{\tau}m).$$

3. Universal norms

Definition 3. *If $K \subset F \subset \mathbf{K}_\infty$, the* **universal norm module** *$U(F)$ is the projective limit*

$$U(F) := \mathbf{Q}_p \otimes \varprojlim_{K \subset L \subset F} (E(L) \otimes \mathbf{Z}_p)$$

(projective limit with respect to traces, over finite *extensions L of K in F) with its natural Λ_F-structure. If F is a finite extension of K, then $U(F)$ is simply $E(F) \otimes \mathbf{Q}_p$.*

If F is a $\mathbf{Z}_p$-extension of K, then $U(F)$ is a free Λ_F-module of finite rank, and is zero if and only if the Mordell-Weil ranks of E over subfields of F are bounded (cf. [8] §18 or [12] §2.2). The first author conjectured some time ago [8] that for $\mathbf{Z}_p$-extensions F/K, and under our running hypotheses, $U(F) = 0$ if $F \neq K_\infty^{\mathrm{anti}}$ and $U(K_\infty^{\mathrm{anti}})$ is free of rank one over Λ_{anti}. The following theorem follows from recent work of Kato [6] for K_∞^{cycl} and Vatsal [15] and Cornut [1] for K_∞^{anti}.

Theorem 4. *$U(K_\infty^{\mathrm{cycl}}) = 0$ and $U(K_\infty^{\mathrm{anti}})$ is free of rank one over Λ_{anti}.*

For the rest of this paper we will write $\mathcal{U}$ for the anticyclotomic universal norm module $U(K_\infty^{\mathrm{anti}})$. Complex conjugation gives $\mathcal{U}$ a natural semi-linear τ-module structure. Since $\mathcal{U}$ is free of rank one over Λ_{anti}, we conclude that $\mathcal{U}$ is completely determined (up to isomorphism preserving its τ-structure) by its sign.

Let $r^{\pm}$ be the rank of the ± 1 eigenspace of τ acting on $E(K)$, so $\operatorname{rank} E(\mathbf{Q}) = r^+$ and $\operatorname{rank} E(K) = r^+ + r^-$. By Proposition 1, $\operatorname{rank} E(K)$ is odd so $r^+ \neq r^-$.

Conjecture 5 (Sign Conjecture). *The sign of the semi-linear τ-module $\mathcal{U}$ is $+1$ if $r^+ > r^-$, and is -1 if $r^- > r^+$.*

Remark. Equivalently, the Sign Conjecture asserts that the sign of $\mathcal{U}$ is $+1$ if twice $\operatorname{rank} E(\mathbf{Q})$ is greater than $\operatorname{rank} E(K)$, and -1 otherwise.

As we discuss below in §4, the Sign Conjecture is related to the nondegeneracy of the p-adic height pairing (see the remark after Conjecture 11).

The Λ_{anti}-module $\mathcal{U}$ comes with a canonical Hermitian structure. That is, the canonical (cyclotomic) p-adic height pairing (see [10] and [12] §2.3)

$$\mathrm{h} : \mathcal{U} \otimes_{\Lambda_{\mathrm{anti}}} \mathcal{U}^{(\tau)} \longrightarrow \Gamma_{\mathrm{cycl}} \otimes_{\mathbf{Z}_p} \Lambda_{\mathrm{anti}}$$

is a τ-Hermitian pairing in the sense of Definition 2.

Conjecture 6 (Height Conjecture). *The homomorphism* h *is an isomorphism of free Λ_{anti}-modules of rank one*

$$\mathrm{h} : \mathcal{U} \otimes_{\Lambda_{\mathrm{anti}}} \mathcal{U}^{(\tau)} \xrightarrow{\sim} \Gamma_{\mathrm{cycl}} \otimes_{\mathbf{Z}_p} \Lambda_{\mathrm{anti}}.$$

The Λ_{anti}-module $\mathcal{U}$ has an important submodule, the **Heegner submodule** $\mathcal{H} \subset \mathcal{U}$. Fix a modular parameterization $X_0(N) \to E$. The Heegner submodule $\mathcal{H}$ is the cyclic Λ_{anti}-module generated by a trace-compatible sequence $c = \{c_L\}$ of Heegner points $c_L \in E(L) \otimes \mathbf{Z}_p$ for finite extensions L of K in K_∞^{anti}. See for example [8] §19 or [12] §3. Call such a $c \in \mathcal{H}$ a **Heegner generator**. The Heegner generators of $\mathcal{H}$ are well-defined up to multiplication by an element of $\pm\Gamma \subset (\Lambda_{\mathrm{anti}})^\times$. The Λ_{anti}-submodule $\mathcal{H} \subset \mathcal{U}$ is stable under the semi-linear τ-structure of $\mathcal{U}$, so the action of τ gives an isomorphism $\mathcal{U}/\mathcal{H} \xrightarrow{\sim} (\mathcal{U}/\mathcal{H})^{(\tau)} \cong \mathcal{U}^{(\tau)}/\mathcal{H}^{(\tau)}$.

Let $c^{(\tau)}$ denote the element c viewed in the Λ_{anti}-module $\mathcal{H}^{(\tau)}$. Since

$$(\pm\gamma c) \otimes_{\Lambda_{\mathrm{anti}}} (\pm\gamma c)^{(\tau)} = c \otimes_{\Lambda_{\mathrm{anti}}} c^{(\tau)}$$

for every $\pm\gamma \in \pm\Gamma$, the element $c \otimes c^{(\tau)} \in \mathcal{H} \otimes_{\Lambda_{\text{anti}}} \mathcal{H}^{(\tau)}$ is independent of the choice of Heegner generator, and is therefore a totally canonical generator of the free, rank one Λ_{anti}-module $\mathcal{H} \otimes_{\Lambda_{\text{anti}}} \mathcal{H}^{(\tau)}$.

Definition 7. *The* **Heegner** L**-function** *(for the triple (E, K, p) satisfying our running hypotheses) is the element*

$$\mathcal{L} := \mathrm{h}(c \otimes c^{(\tau)}) \in \Gamma_{\text{cycl}} \otimes_{\mathbf{Z}_p} \Lambda_{\text{anti}}.$$

Conjecture 8. $\Gamma_{\text{cycl}} \otimes \mathrm{char}(\mathcal{U}/\mathcal{H})^2 = \Lambda_{\text{anti}}\mathcal{L}$ *inside* $\Gamma_{\text{cycl}} \otimes \Lambda_{\text{anti}}$.

One sees easily that $\Gamma_{\text{cycl}} \otimes \mathrm{char}(\mathcal{U}/\mathcal{H})^2 \supset \Lambda_{\text{anti}}\mathcal{L}$, and that Conjecture 8 is equivalent to the Height Conjecture (Conjecture 6).

4. The analytic theory

The ("two-variable") p-adic L-function for E over K is an element $\mathbf{L} \in \mathbf{\Lambda}$ constructed by Haran [3] and by a different, more general, method by Hida [4] (see also the papers of Perrin-Riou [13, 14]). The L-function $\mathbf{L}$ is characterized by the fact that it interpolates special values of the classical Hasse-Weil L-function of twists of E over K. More precisely, embedding $\bar{\mathbf{Q}}$ both in $\mathbf{C}$ and $\bar{\mathbf{Q}}_p$, if $\chi : \mathbf{\Gamma} \to \bar{\mathbf{Z}}^\times \subset \bar{\mathbf{Z}}_p^\times$ is a character of finite order then

$$\chi(\mathbf{L}) = c(\chi)\frac{L_{\text{classical}}(E_{/K}, \chi, 1)}{8\pi^2 \|f_E\|^2} \tag{4.1}$$

where $L_{\text{classical}}(E_{/K}, \chi, s)$ is the Hasse-Weil L-function of the twist of $E_{/K}$ by χ, $c(\chi)$ is an explicit algebraic number (cf. [13] Théorème 1.1), f_E is the modular form on $\Gamma_0(N)$ corresponding to E, and $\|f_E\|$ is its Petersson norm.

Projecting $\mathbf{L} \in \mathbf{\Lambda}$ to the cyclotomic or the anticyclotomic line via the natural projections $\mathbf{\Lambda} \to \Lambda_{\text{cycl}}$ and $\mathbf{\Lambda} \to \Lambda_{\text{anti}}$, we get "one-variable" p-adic L-functions

$$\mathbf{L} \mapsto L_{\text{cycl}} \in \Lambda_{\text{cycl}} \qquad \text{and} \qquad \mathbf{L} \mapsto L_{\text{anti}} \in \Lambda_{\text{anti}}.$$

It follows from the functional equation satisfied by $\mathbf{L}$ ([13] Théorème 1.1) and the Heegner Hypothesis that $L_{\text{anti}} = 0$. In other words, viewing $\mathbf{\Lambda} = \Lambda_{\text{anti}}[[\Gamma_{\text{cycl}}]]$ as the completed group ring of Γ_{cycl} with coefficients in Λ_{anti}, we have that the "constant term" of $\mathbf{L} \in \Lambda_{\text{anti}}[[\Gamma_{\text{cycl}}]]$ vanishes. We now consider its "linear term."

There is a canonical isomorphism of (free, rank one) Λ_{anti}-modules

$$\Gamma_{\text{cycl}} \otimes_{\mathbf{Z}_p} \Lambda_{\text{anti}} \cong \mathbf{I}_{K_\infty^{\text{anti}}}/\mathbf{I}_{K_\infty^{\text{anti}}}^2$$

which sends $\gamma \otimes 1 \in \Gamma_{\text{cycl}} \otimes_{\mathbf{Z}_p} \Lambda_{\text{anti}}$ to $\gamma - 1 \in \mathbf{I}_{K_\infty^{\text{anti}}}/\mathbf{I}_{K_\infty^{\text{anti}}}^2$.

Conjecture 9 (Λ-adic Gross-Zagier Conjecture). *Let L' denote the image of $\mathbf{L}$ under the map $\mathbf{I}_{K_\infty^{\text{anti}}}/\mathbf{I}_{K_\infty^{\text{anti}}}^2 \xrightarrow{\sim} \Gamma_{\text{cycl}} \otimes_{\mathbf{Z}_p} \Lambda_{\text{anti}}$. Then*

$$L' = d^{-1}\mathcal{L}$$

where d is the degree of the modular parametrization $X_0(N) \to E$.

Remark. Perrin-Riou [13] proved that if p splits in K and the discriminant D of K is odd, then L' and $d^{-1}\mathcal{L}$ have the same image under the projection $\Lambda_{\text{anti}} \to \Lambda_K = \mathbf{Q}_p$.

Let $\mathbf{I} := \mathbf{I}_K$, the augmentation ideal of $\mathbf{\Lambda}$. For every integer $r \geq 0$ we have $\mathbf{I}^r/\mathbf{I}^{r+1} \cong \mathrm{Sym}^r_{\mathbf{Z}_p}(\mathbf{\Gamma}) \otimes \mathbf{Q}_p$. Using the direct sum decomposition $\mathbf{\Gamma} = \Gamma_{\text{cycl}} \oplus \Gamma_{\text{anti}}$ we get a canonical direct sum decomposition

$$\mathrm{Sym}^r_{\mathbf{Z}_p}(\mathbf{\Gamma}) = \bigoplus_{j=0}^{r} \Gamma^{r-j,j} \quad \text{where } \Gamma^{i,j} := (\Gamma_{\text{cycl}})^{\otimes i} \otimes_{\mathbf{Z}_p} (\Gamma_{\text{anti}})^{\otimes j}. \tag{4.2}$$

Consider the canonical (two-variable) p-adic height pairing

$$\langle\ ,\ \rangle\ :\ E(K) \times E(K) \longrightarrow \mathbf{\Gamma} \otimes \mathbf{Q}_p. \tag{4.3}$$

Set $r = \operatorname{rank} E(K)$, which is odd by Proposition 1. Define the **two-variable p-adic regulator** $R_p(E,K)$ to be the discriminant of this pairing:

$$R_p(E,K) := t^{-2} \det\langle P_i, P_j\rangle \in \mathrm{Sym}^r_{\mathbf{Z}_p}(\mathbf{\Gamma}) \otimes \mathbf{Q}_p \cong \mathbf{I}^r/\mathbf{I}^{r+1},$$

where $\{P_1, \dots, P_r\}$ generates a subgroup of $E(K)$ of finite index t. For each integer $j = 0, \dots, r$ let $R_p(E,K)^{r-j,j}$ be the projection of $R_p(E,K)$ into $\Gamma^{r-j,j} \otimes \mathbf{Q}_p$ under (4.2), so that

$$R_p(E,K) = \bigoplus_{j=0}^{r} R_p(E,K)^{r-j,j}.$$

Recall that $r^{\pm}$ is the rank of the ± 1-eigenspace $E(K)^{\pm}$ of τ acting on $E(K)$.

Proposition 10. *$R_p(E,K)^{r-j,j} = 0$ unless j is even and $j \leq 2\min(r^+, r^-)$.*

Proof. This follows from the fact that the height pairing (4.3) is τ-Hermitian, so $\langle \tau x, \tau y\rangle = \langle x, y\rangle^{\tau}$, and therefore the induced height pairings

$$E(K)^{\pm} \times E(K)^{\pm} \to \Gamma_{\text{anti}} \otimes \mathbf{Q}_p, \quad E(K)^{+} \times E(K)^{-} \to \Gamma_{\text{cycl}} \otimes \mathbf{Q}_p$$

vanish.

Conjecture 11 (Maximal nondegeneracy of the height pairing). *If j is even and $0 \leq j \leq 2\min(r^+, r^-)$ then $R_p(E,K)^{r-j,j} \neq 0$.*

Remark. Conjecture 11, or more specifically the nonvanishing of $R_p(E,K)^{r-j,j}$ when $j = 2\min(r^+, r^-)$, implies the Sign Conjecture (Conjecture 5). This is proved in the same way as Proposition 10, using the additional fact that the anticyclotomic universal norms in $E(K) \otimes \mathbf{Z}_p$ are in the kernel of the anticyclotomic p-adic height pairing $(E(K) \otimes \mathbf{Z}_p) \times (E(K) \otimes \mathbf{Z}_p) \to \Gamma_{\text{anti}} \otimes \mathbf{Q}_p$.

5. The arithmetic theory

For every algebraic extension F of K, let $\mathrm{Sel}_p(E_{/F})$ denote the p-power Selmer group of E over F, the subgroup of $H^1(G_F, E[p^\infty])$ that sits in an exact sequence

$$0 \longrightarrow E(F) \otimes \mathbf{Q}_p/\mathbf{Z}_p \longrightarrow \mathrm{Sel}_p(E_{/F}) \longrightarrow Ш(E_{/F})[p^\infty] \longrightarrow 0$$

where $Ш(E_{/F})$ is the Shafarevich-Tate group of E over F. Also write

$$\mathcal{S}_p(E_{/F}) = \mathrm{Hom}(\mathrm{Sel}_p(E_{/F}), \mathbf{Q}_p/\mathbf{Z}_p) \otimes \mathbf{Q}_p$$

for the tensor product of $\mathbf{Q}_p$ with the Pontrjagin dual of the Selmer group.

The following theorem is proved using techniques which go back to [7]; see [2] and [12] Lemme 5, §2.2.

Theorem 12 (Control Theorem). *Suppose* $K \subset F \subset \mathbf{K}_\infty$.

(i) *The natural restriction map* $H^1(F, E[p^\infty]) \to H^1(\mathbf{K}_\infty, E[p^\infty])$ *induces an isomorphism* $\mathcal{S}_p(E_{/\mathbf{K}_\infty}) \otimes_{\mathbf{\Lambda}} \Lambda_F \xrightarrow{\sim} \mathcal{S}_p(E_{/F})$.

(ii) *There is a canonical isomorphism* $U(F) \xrightarrow{\sim} \mathrm{Hom}_{\Lambda_F}(\mathcal{S}_p(E_{/F}), \Lambda_F)$.

Conjecture 13 (Two-variable main conjecture [8, 12]). *The two-variable p-adic L-function* $\mathbf{L}$ *generates the ideal* $\mathrm{char}_{\mathbf{\Lambda}}(\mathcal{S}_p(E_{/\mathbf{K}_\infty}))$ *of* $\mathbf{\Lambda}$.

Restricting the two-variable main conjecture to the cyclotomic and anticyclotomic lines leads to the following "one-variable" conjectures originally formulated in [9] and [12], respectively. Let L' denote the image of $\mathbf{L}$ in $\Gamma_{\mathrm{cycl}} \otimes_{\mathbf{Z}_p} \Lambda_{\mathrm{anti}}$ as in Conjecture 9, and $\mathcal{S}_p(E_{/K_\infty^{\mathrm{anti}}})_{\mathrm{tors}}$ the Λ_{anti}-torsion submodule of $\mathcal{S}_p(E_{/K_\infty^{\mathrm{anti}}})$.

Conjecture 14 (Cyclotomic and anticyclotomic main conjectures).

(i) L_{cycl} *generates the ideal* $\mathrm{char}_{\Lambda_{\mathrm{cycl}}}(\mathcal{S}_p(E_{/K_\infty^{\mathrm{cycl}}}))$ *of* Λ_{cycl}.

(ii) L' *generates* $\Gamma_{\mathrm{cycl}} \otimes \mathrm{char}_{\Lambda_{\mathrm{anti}}}(\mathcal{S}_p(E_{/K_\infty^{\mathrm{anti}}})_{\mathrm{tors}})$ *inside* $\Gamma_{\mathrm{cycl}} \otimes \Lambda_{\mathrm{anti}}$.

Remark. Using Euler systems, Kato [6] and Howard [5] have proved (under some mild additional hypotheses) divisibilities related to the cyclotomic and anticyclotomic main conjectures, respectively, namely

$$L_{\mathrm{cycl}} \in \mathrm{char}_{\Lambda_{\mathrm{cycl}}}(\mathcal{S}_p(E_{/K_\infty^{\mathrm{cycl}}})), \quad \mathrm{char}_{\Lambda_{\mathrm{anti}}}(\mathcal{U}/\mathcal{H})^2 \subset \mathrm{char}_{\Lambda_{\mathrm{anti}}}(\mathcal{S}_p(E_{/K_\infty^{\mathrm{cycl}}})_{\mathrm{tors}})$$

(note that Conjectures 8 and 9 predict that $\Gamma_{\mathrm{cycl}} \otimes \mathrm{char}_{\Lambda_{\mathrm{anti}}}(\mathcal{U}/\mathcal{H})^2 = L'\Lambda_{\mathrm{anti}}$).

Conjecture 15 (Two-variable p-adic BSD conjecture). *Let* $r = \mathrm{rank}\, E(K)$. *The two-variable p-adic L-function* $\mathbf{L} \in \mathbf{\Lambda}$ *is contained* $\mathbf{I}^r$ *and*

$$\mathbf{L} \equiv c(\chi_{\mathrm{triv}}) \#(Ш(E_{/K})) \prod_v c_v \cdot R_p(E, K) \pmod{\mathbf{I}^{r+1}}$$

where $c(\chi_{\mathrm{triv}})$ *is the rational number in the interpolation formula* (4.1) *for the trivial character,* $Ш(E_{/K})$ *is the Shafarevich-Tate group of* E *over* K, *and the* c_v *are the Tamagawa factors in the (usual) Birch and Swinnerton-Dyer conjecture for* E *over* K.

6. Orthogonal Λ-modules

In this final section we introduce a purely algebraic template which, when it "fits", gives rise to many of the properties conjectured in the previous sections.

Keep the notation of the previous sections. In particular $\tau : \mathbf{\Lambda} \to \mathbf{\Lambda}$ is the involution of $\mathbf{\Lambda}$ induced by complex conjugation on K, and if V is a $\mathbf{\Lambda}$-module, then $V^{(\tau)}$ denotes V with $\mathbf{\Lambda}$-module structure obtained by composition with τ. Let $V^* = \mathrm{Hom}_{\mathbf{\Lambda}}(V, \mathbf{\Lambda})$. If V is a free $\mathbf{\Lambda}$-module of rank r, then $\det_{\mathbf{\Lambda}}(V^\tau)$ will denote the r-th exterior power of V and a **τ-gauge** on V is a $\mathbf{\Lambda}$-isomorphism between the free $\mathbf{\Lambda}$-modules of rank one

$$t_V \ : \ \det_{\mathbf{\Lambda}}(V^*) \cong \det_{\mathbf{\Lambda}}(V^{(\tau)})$$

or equivalently an isomorphism $\det_{\mathbf{\Lambda}}(V) \otimes \det_{\mathbf{\Lambda}}(V^{(\tau)}) \xrightarrow{\sim} \mathbf{\Lambda}$.

By an **orthogonal $\mathbf{\Lambda}$-module** we mean a free $\mathbf{\Lambda}$-module V with semi-linear τ-structure endowed with a τ-gauge t_V and a $\mathbf{\Lambda}$-bilinear τ-Hermitian pairing (Definition 2)

$$\pi : V \otimes_{\mathbf{\Lambda}} V^{(\tau)} \longrightarrow \mathbf{\Lambda}.$$

Viewing π as a $\mathbf{\Lambda}$-linear map $V^{(\tau)} \to V^*$, the composition

$$t_V \circ \det_{\mathbf{\Lambda}}(\pi) \ : \ \det_{\mathbf{\Lambda}}(V^{(\tau)}) \longrightarrow \det_{\mathbf{\Lambda}}(V^*) \longrightarrow \det_{\mathbf{\Lambda}}(V^{(\tau)})$$

must be multiplication by an element $\mathrm{disc}(V) \in \mathbf{\Lambda}$ that we call the **discriminant** of the orthogonal $\mathbf{\Lambda}$-module V. We further assume that $\mathrm{disc}(V) \neq 0$, and we define $M = M(V, \pi)$ to be the cokernel of the (injective) map $\pi : V^{(\tau)} \to V^*$, so we have

$$0 \longrightarrow V^{(\tau)} \longrightarrow V^* \longrightarrow M \longrightarrow 0. \tag{6.1}$$

If $K \subset F \subset \mathbf{K}_\infty$, recall that $\mathbf{I}_F = \ker\{\mathbf{\Lambda} \twoheadrightarrow \Lambda_F\}$ and define

$$V(F) := \{x \in V : \pi(x, V^{(\tau)}) \subset \mathbf{I}_F\}/\mathbf{I}_F V = \ker\{V \otimes_{\mathbf{\Lambda}} \Lambda_F \xrightarrow{\pi \otimes 1} (V^{(\tau)})^* \otimes_{\mathbf{\Lambda}} \Lambda_F\}$$

and similarly $V^{(\tau)}(F) := \ker\{V^{(\tau)} \otimes \Lambda_F \to V^* \otimes \Lambda_F\}$. Any lift $\tilde{\tau}$ of τ to $\mathrm{Gal}(\mathbf{K}_\infty/\mathbf{Q})$ induces an isomorphism $V(F) \to V^{(\tau)}(F)$. From (6.1) we obtain

$$0 \longrightarrow V^{(\tau)}(F) \longrightarrow V^{(\tau)} \otimes_{\mathbf{\Lambda}} \Lambda_F \longrightarrow V^* \otimes_{\mathbf{\Lambda}} \Lambda_F \longrightarrow M \otimes_{\mathbf{\Lambda}} \Lambda_F \longrightarrow 0 \tag{6.2}$$

and (applying $\mathrm{Hom}(\,\cdot\,, \Lambda_F)$ and using the Hermitian property of π)

$$V(F) \cong \mathrm{Hom}_{\Lambda_F}(M \otimes_{\mathbf{\Lambda}} \Lambda_F, \Lambda_F). \tag{6.3}$$

We have an induced pairing

$$\pi_F : V^{(\tau)}(F) \otimes_{\Lambda_F} V(F) \longrightarrow \mathbf{I}_F/\mathbf{I}_F^2,$$

which we call the F-derived pairing. If F is stable under complex conjugation then $V^{(\tau)}(F)$ is canonically isomorphic to $V(F)^{(\tau)}$ and π_F is τ-Hermitian.

Now suppose $F = K_\infty^{\text{anti}}$. By (6.3), $V(K_\infty^{\text{anti}})$ is free over Λ_{anti}. Applying the determinant functor to (6.2), the τ-gauge t_V induces an isomorphism

$$\det_{\Lambda_{\text{anti}}} V(K_\infty^{\text{anti}})^{(\tau)} = \det_{\Lambda_{\text{anti}}} V^{(\tau)}(K_\infty^{\text{anti}}) \xrightarrow{\sim} \text{Hom}(\det_{\Lambda_{\text{anti}}}(M \otimes_{\mathbf{\Lambda}} \Lambda_{\text{anti}}), \Lambda_{\text{anti}}).$$

If $V(K_\infty^{\text{anti}})$ has rank one over Λ_{anti}, then $V(K_\infty^{\text{anti}})$ contains a unique maximal τ-stable submodule H such that the map

$$V(K_\infty^{\text{anti}})^{(\tau)} \xrightarrow{\sim} \text{Hom}(\det_{\Lambda_{\text{anti}}}(M \otimes \Lambda_{\text{anti}}), \Lambda_{\text{anti}}) \supset \text{Hom}(M \otimes \Lambda_{\text{anti}}, \Lambda_{\text{anti}}) \cong V(K_\infty^{\text{anti}})$$

sends $H^{(\tau)}$ into H. (Namely, $H = JV(K_\infty^{\text{anti}})$ where J is the largest ideal of Λ_{anti} such that $J^\tau = J$ and $J^2 \subset \text{char}_{\Lambda_{\text{anti}}}(M \otimes \Lambda_{\text{anti}})_{\text{tors}}$.)

Recall that $\text{Sel}_p(E_{/F})$ denotes the p-power Selmer group of E over F and $\mathcal{S}_p(E_{/F}) = \text{Hom}(\text{Sel}_p(E_{/F}), \mathbf{Q}_p/\mathbf{Z}_p) \otimes \mathbf{Q}_p$.

Proposition 16. *With notation as above, suppose that V is an orthogonal $\mathbf{\Lambda}$-module and $\varphi_V : M \xrightarrow{\sim} \mathcal{S}_p(E_{/\mathbf{K}_\infty})$ is an isomorphism. Then for every extension F of K in $\mathbf{K}_\infty$, φ_V induces an isomorphism*

$$V(F) \xrightarrow{\sim} U(F)$$

where $U(F)$ is the universal norm module defined in §3.

proof. This follows directly from Theorem 12 and (6.3).

Definition 17. *We say that the orthogonal $\mathbf{\Lambda}$-module V* **organizes the anticyclotomic arithmetic of** *(E, K, p) if the following three properties hold.*

(a) $\text{disc}(V) = \mathbf{L}$, *the two-variable p-adic L-function of E.*
(b) There is an isomorphism $\varphi_V : M \xrightarrow{\sim} \mathcal{S}_p(E_{/\mathbf{K}_\infty})$.
(c) The isomorphism $V(K_\infty^{\text{anti}}) \cong \mathcal{U}$ of Proposition 16 identifies $H \subset V(K_\infty^{\text{anti}})$ with the Heegner submodule $\mathcal{H} \subset \mathcal{U}$, and identifies the K_∞^{anti}-derived pairing with the canonical p-adic height pairing into $\mathbf{I}_{K_\infty^{\text{anti}}}/\mathbf{I}^2_{K_\infty^{\text{anti}}} \cong \Gamma_{\text{cycl}} \otimes \Lambda_{\text{anti}}$.

Question. Given E, K, and p satisfying our running hypotheses, is there an orthogonal $\mathbf{\Lambda}$-module V that organizes the anticyclotomic arithmetic of (E, K, p)?

If one is not quite so (resp., much more) optimistic one could formulate an analogous question with the ring $\mathbf{\Lambda}$ replaced by the localization of $\mathbf{\Lambda}$ at $\mathbf{I}$ (resp., with $\mathbf{\Lambda}$ replaced by $\mathbf{Z}_p[[\mathbf{\Gamma}]]$).

Question. If V is an orthogonal $\mathbf{\Lambda}$-module V which organizes the anticyclotomic arithmetic of (E, K, p), then for every finite extension F of K in $\mathbf{K}_\infty$, we have an isomorphism $E(F) \otimes \mathbf{Q}_p = U(F) \cong V(F)$ as in Proposition 16, a p-adic height pairing on $E(F) \otimes \mathbf{Q}_p$, and the F-derived pairing on $V(F)$. How are these pairings related?

When $F = K_\infty^{\text{anti}}$ condition (c) says that the two pairings are the same, but it seems that in general they cannot be the same for finite extensions F/K.

Theorem 18. *Suppose that there is an orthogonal $\mathbf{\Lambda}$-module V that organizes the anticyclotomic arithmetic of (E, K, p). Then Conjectures 13 (the 2-variable main conjecture), and 14(i) (the cyclotomic main conjecture) hold.*

If further the induced pairing $V(K_\infty^{\mathrm{anti}}) \otimes V(K_\infty^{\mathrm{anti}})^{(\tau)} \to \Gamma_{\mathrm{cycl}} \otimes \Lambda_{\mathrm{anti}}$ is surjective, then Conjectures 6 (the Height Conjecture), 8, 9 (the Λ-adic Gross-Zagier conjecture), and 14(ii) (the anticyclotomic main conjecture) also hold.

Brief outline of the proof of Theorem 18. Since $\mathrm{disc}(V)$ is a generator of $\mathrm{char}_{\mathbf{\Lambda}}(M)$, the two-variable main conjecture follows immediately from (a) and (b) of Definition 17. The cyclotomic main conjecture follows from the two-variable main conjecture.

Now suppose that the induced pairing $V(K_\infty^{\mathrm{anti}}) \otimes V(K_\infty^{\mathrm{anti}})^{(\tau)} \to \Gamma_{\mathrm{cycl}} \otimes \Lambda_{\mathrm{anti}}$ is surjective. By (c) of Definition 17 this is equivalent to the Height Conjecture, which in turn is equivalent to Conjecture 8.

Howard proved in [5] that $\mathcal{S}_p(E_{/K_\infty^{\mathrm{anti}}})$ is pseudo-isomorphic to $\Lambda_{\mathrm{anti}} \oplus B^2$ where B is a τ-stable torsion Λ_{anti}-module. By Theorem 12(i) the same is true of $M \otimes \Lambda_{\mathrm{anti}}$, and so the remark at the end of the definition of H shows that $H = \mathrm{char}(B)V(K_\infty^{\mathrm{anti}})$. Using (6.2), (6.3), and our assumption that the induced pairing is surjective, one can show that the image of $\mathbf{L}$ in $\mathbf{I}_{K_\infty^{\mathrm{anti}}}/\mathbf{I}_{K_\infty^{\mathrm{anti}}}^2$ generates $\mathrm{char}(B)^2\mathbf{I}_{K_\infty^{\mathrm{anti}}}/\mathbf{I}_{K_\infty^{\mathrm{anti}}}^2$. The Λ-adic Gross-Zagier conjecture and the anticyclotomic main conjecture follow from these facts and (c). □

References

[1] C. Cornut, Mazur's conjecture on higher Heegner points, *Invent. math.* **148** (2002), 495–523.

[2] R. Greenberg, Galois theory for the Selmer group of an abelian variety (preprint).

[3] S. Haran, p-adic L-functions for elliptic curves over CM fields, thesis, MIT 1983.

[4] H. Hida, A p-adic measure attached to the zeta functions associated with two elliptic modular forms. I, *Invent. Math.* **79** (1985), 159–195.

[5] B. Howard, *The Heegner point Kolyvagin system*, thesis, Stanford University 2002.

[6] K. Kato, p-adic Hodge theory and values of zeta functions of modular forms (preprint).

[7] B. Mazur, Rational points of abelian varieties with values in towers of number fields, *Invent. Math.* **18** (1972), 183–266.

[8] ———, Modular curves and arithmetic. In: *Proceedings of the International Congress of Mathematicians (Warsaw, 1983)*, PWN, Warsaw (1984), 185–211.

[9] B. Mazur, P. Swinnerton-Dyer, Arithmetic of Weil curves, *Invent. Math.* **25** (1974), 1–61.

[10] B. Mazur, J. Tate, Canonical height pairings via biextensions. In: *Arithmetic and Geometry*, Progr. Math. **35**, Birkhaüser, Boston (1983), 195–237.

[11] J. Nekovář, On the parity of ranks of Selmer groups. II, *C. R. Acad. Sci. Paris Sér. I Math.* **332** (2001), 99–104.

[12] B. Perrin-Riou, Fonctions L p-adiques, théorie d'Iwasawa et points de Heegner, *Bull. Soc. Math. France* **115** (1987), 399–456.
[13] ———, Points de Heegner et dérivées de fonctions L p-adiques, *Invent. Math.* **89** (1987), 455–510.
[14] ———, Fonctions L p-adiques associées à une forme modulaire et à un corps quadratique imaginaire, *J. London Math. Soc.* **38** (1988), 1–32.
[15] V. Vatsal, Special values of anticyclotomic L-functions (preprint).

ICM 2002 · Vol. II · 197–206

Théorie Ergodique et Géométrie Arithmétique

Emmanuel Ullmo*

Abstract

We will present several examples in which ideas from ergodic theory can be useful to study some problems in arithmetic and algebraic geometry.

2000 Mathematics Subject Classification: 11F32, 11G10, 11G15, 11G40, 22D40, 22E40.
Keywords and Phrases: Equidistribution, Variétés abeliennes, Variétés de Shimura.

1. Introduction

Le but de ce rapport est d'expliquer différentes techniques permettant de montrer l'équidistribution de certains ensembles de points de nature arithmétique sur des variétés algébriques définies sur des corps de nombres et de donner des applications arithmétiques et géométriques de ces résultats.

Si X est une variété algébrique sur $\mathbb{C}$ et E une ensemble fini de $X(\mathbb{C})$ on note $|E|$ son cardinal et Δ_E la mesure de Dirac normalisée

$$\Delta_E = \frac{1}{|E|} \sum_{x \in E} \delta_x.$$

Si E_n est une suite d'ensembles finis de $X(\mathbb{C})$ et μ une mesure de probabilité sur $X(\mathbb{C})$, on dit que les E_n sont équidistribués pour μ si pour toute fonction continue bornée f sur $X(\mathbb{C})$ on a

$$\Delta_{E_n}(f) = \frac{1}{|E_n|} \sum_{x \in E_n} f(x) \longrightarrow \int_{X(\mathbb{C})} f\mu.$$

*Universit Paris-Sud Orsay Bât 425, 91405 Orsay Cedex France. E-mail: ullmo@math.u-psud.fr

Soit X une variété algébrique, une suite de points x_n de X est dite "générique" si pour toute sous-variété Y de X, $Y \neq X$, $\{n \in \mathbb{N}\ , x_n \in Y\}$ est un ensemble fini. (Il revient au même de dire que x_n converge vers le point générique pour la topologie de Zariski).

André et Oort ont formulé un analogue de la conjecture de Manin-Mumford démontrée par Raynaud [18] [19] dans le cadre des variétés de Shimura. Dans ces deux conjectures, on dispose de points spéciaux et de variétés spéciales. Pour la conjecture de Manin-Mumford l'espace ambiant est une variété abélienne, les points spéciaux sont les points de torsion et les variétés spéciales sont les "sous-variétés de torsion" (translatés, par un point de torsion, d'une sous-variété abélienne). Pour la conjecture d'André-Oort l'espace ambiant est une variété de Shimura, les points spéciaux sont les points à multiplication complexe (ou points CM) et les sous-variétés spéciales sont les "sous-variétés de type de Hodge" (des composantes irréductibles de translatés par un opérateur de Hecke de sous-variétés de Shimura). Nous préciserons ces définitions plus bas. Dans les deux cas ces conjectures s'énoncent sous la forme: une composante irréductible de l'adhérence de Zariski d'un ensemble de points spéciaux est une sous-variété spéciale.

Dans ce cadre une suite de points x_n de X (X variété abélienne ou X variété de Shimura) est dite "stricte" si pour toute sous-variété spéciale Y de X, $Y \neq X$, $\{n \in \mathbb{N}\ , x_n \in Y\}$ est un ensemble fini. On remarque qu'avec ces définitions les conjectures d'André-Oort et de Manin-Mumford se retraduisent de la manière suivante: Toute suite stricte de points spéciaux est générique.

Une conséquence géométrique (conjecturale pour les variétés de Shimura) que l'on obtient en considérant l'adhérence de Zariski de l'ensemble des points spéciaux d'une sous-variété M de X est l'existence d'un ensemble fini $\{S_1, \ldots, S_r\}$ de sous-variétés spéciales avec $S_i \subset M$ telle que toute sous-variété spéciale $S \subset M$ est contenue dans l'un des S_i.

Dans la première partie nous décrivons des résultats d'équidistribution pour des suites de points de petite hauteur sur des variétés algébriques utilisant la géométrie d'Arakelov. Le résultat le plus marquant est la résolution de la conjecture de Bogomolov (qui généralise la conjecture de Manin-Mumford et en donne une nouvelle démonstration) pour les variétés abéliennes due à Zhang [24] et à l'auteur du rapport [22].

Dans la deuxième partie nous expliquons des résutats d'équidistribution de points de Hecke sur des variétés de la forme $X = \Gamma \backslash G(\mathbb{R})$ pour un groupe algébrique semi-simple et simplement connexe G et un réseau Γ. Les méthodes combinent théorie spectrale et théorie des représentations.

Dans la troisième partie nous présentons des énoncés largement conjecturaux pour l'équidistribution des points à multiplication complexe des variétés de Shimura. La théorie analytique des nombres via les familles de fonctions L et la théorie des formes automorphes y jouent un rôle central.

Dans une dernière partie nous expliquons comment la théorie de Ratner et Margulis permet de démontrer des résultats d'équidistribution pour des suites de sous-variétés "fortement spéciales" (appartenant à une classe assez large de sous-variétés spéciales de dimension positive) des variétés de Shimura. Nous expli-

querons la relation avec la conséquence géométrique de la conjecture d'André-Oort précédemment décrite.

2. Equidistribution des points de petite hauteur

Exemple 2.1 On prend $X = \mathbb{G}_m$, E_n l'ensemble des racines n-ième de l'unité, E_n est équidistribué pour la mesure uniforme sur le cercle unité $\frac{d\alpha}{2\pi}$. En utilisant l'irréductibilité du polynome cyclotomique on voit que l'orbite sous Galois d'une racine n-ième primitive de l'unité est aussi équidistribuée pour $\frac{d\alpha}{2\pi}$.

Exemple 2.2 On prend $X = E$ une courbe elliptique sur $\mathbb{C}$ et E_n l'ensemble des points de n torsion, alors E_n est équidistribué pour la mesure de Harr normalisée sur $E(\mathbb{C})$. Si E est défini sur un coprs de nombres K et E n'a pas de multiplication complexe, par le théorème de l'image ouverte de Serre, pour tout nombre premier p assez grand le groupe de Galois agit transitivement sur les points d'ordre p. On en déduit encore que les orbites sous Galois des points d'ordre p sont équidistribuées pour la mesure de Haar normalisée.

La théorie d'Arakelov a permis de comprendre ces énoncés d'une manière bien plus générale. On montre [21] pour une variété arithmétique un théorème général d'équidistribution des orbites sous Galois de suite génériques de points dont la hauteur (à la Arakelov) tend vers 0. Les exemples précédents correspondent à des suites de points de hauteurs nulles. Pour les variétés abéliennes on obtient avec Szpiro et Zhang le résultat suivant (qui donne des informations nouvelles même pour les points de torsion des courbes elliptiques à multiplication complexe):

Théorème 2.3 [21] *Soit A une variété abélienne sur un corps de nombres K. On note h_{NT} la hauteur de Néron-Tate sur les points algébriques de A (associée à un fibré inversible ample symétrique sur X). Soit x_n une suite générique de points algébriques de A telle que $h_{NT}(x_n)$ tend vers 0. Pour toute place à l'infini σ l'orbite sous Galois de x_n est équidistribuée pour la mesure de Haar normalisée $d\mu_\sigma$ de $A_\sigma(\mathbb{C})$.*

L'analogue de cet énoncé pour $\mathbb{G}_m^r$ a été montré par Bilu [2] sans théorie d'Arakelov. Une extension pour certaines variétés semi-abéliennes de ces résultats a été obtenue par Chambert-Loir [6] par des méthodes Arakeloviennes. On peut aussi comprendre grâce aux travaux de Autissier [1] l'exemple 2.1 comme un cas particulier de théorème d'équidistribution vers la mesure d'équilibre d'un compact de capacité 1 de l'orbite sous Galois d'une suite de points entiers algébriques.

On trouvera dans [25] comment on obtient la conjecture de Bogomolov en produisant une contradiction sur les mesures limites de suites de mesures associées à des orbites sous Galois de points de petite hauteur. Retenons l'énoncé suivant dû à l'auteur [22] pour les courbes de genre $g \geq 2$ dans leur jacobienne et étendu en dimension arbitraire par Zhang [24]:

Théorème 2.4 *Soit X une sous-variété d'une variété abélienne A définie sur un corps de nombres K. Grâce à la conjecture de Manin-Mumford démontrée par Raynaud [19], on sait qu'il existe des sous-variétés de torsion (éventuellement de dimension 0) $\{T_1, \ldots, T_r\}$, $T_i \subset X$ tels que si $T \subset X$ est une sous-variété de torsion alors $T \subset T_i$ pour un certain i. Il existe alors $c > 0$ tel que si P est un point algébrique de X et $P \notin \cup_{i=1}^r T_i$ alors $h_{NT}(P) \geq c$.*

3. Equidistribution des points de Hecke

Soient G un groupe algébrique linéaire presque simple et simplement connexe sur $\mathbb{Q}$, $\Gamma \subset G(\mathbb{Q})$ un réseau de congruence et $X = \Gamma\backslash G(\mathbb{R})$. Soit μ_0 la mesure invariante normalisée sur X. Pour tout $a \in G(\mathbb{Q})$ on a une décomposition

$$\Gamma a \Gamma = \cup_{i=1}^{deg(a)} \Gamma a_i$$

avec $deg(a) = |\Gamma\backslash\Gamma a\Gamma| \in \mathbb{N}$. Pour tout $x \in X$, on note $T_a.x$ l'ensemble des $a_i x$ compté avec multiplicité. L'opérateur de Hecke T_a ainsi défini est une correspondance de degré $deg(a)$ sur X; il induit une opération sur les espaces de fonctions $L^2(X, \mu_0)$ (fonctions de carrés intégrables sur X) et $C_b^0(X)$ (fonctions continues bornées sur X) par

$$T_a.f(x) = \frac{1}{deg(a)} \sum_{i=1}^{deg(a)} f(a_i x).$$

Avec Clozel et Oh nous obtenons [3]:

Théorème 3.1 *On suppose que le $\mathbb{Q}$-rang de G est différent de 0. Soit $a_n \in G(\mathbb{Q})$ une suite telle que $deg(a_n) \to \infty$. Pour tout $x \in X$ les $T_{a_n}.x$ sont équidistribués pour μ_0. De plus pour tout $f \in L^2(X, \mu_0)$ on a la convergence L^2*

$$\|T_{a_n} f - \int_X f.\mu_0\|_{L^2} \longrightarrow 0.$$

On a en fait des résultats aussi dans le cas ou le $\mathbb{Q}$-rang de G vaut 0. La méthode de démonstration fournit des estimations très précises pour la vitesse de convergence dans le théorème L^2. Si on dispose de plus de régularité sur f (par exemple f C^∞ à support compact), cete vitesse est obtenue aussi pour la convergence simple (ou uniforme sur les compacts). Pour $G = SL_n$ ($n \geq 3$) ou $G = Sp_{2n}$ ($n \geq 2$) ces estimations sont essentiellement optimales.

On montre par des méthodes classiques que l'énoncé de convergence simple du théorème se déduit de l'énoncé L^2. Pour montrer le théorème L^2 on écrit la décomposition spectrale de $L^2(X, \mu_0)$ sous la forme adélique. Une fonction ϕ intervenant dans la décomposition spectrale est alors propre pour les opérateurs de Hecke et les valeurs propres s'interprètent comme des coefficients matriciaux de représentations locales associées à ϕ. Pour montrer le théorème sous la forme L^2, on doit montrer que $T_{a_n}\phi \to 0$ quand $n \to \infty$ au sens L^2. On se ramène ainsi à contrôler la décroissance de ces coefficients matriciaux. En $\mathbb{Q}$-rang $r \geq 2$ on dispose

d'assez d'informations sur le dual unitaire pour conclure grâce aux travaux de Oh ([17], théorème 5.7). En $\mathbb{Q}$-rang 1 on utilise un principe de restriction à la Burger-Sarnak en une place finie démontré dans [4] et une approximation de la conjecture de Ramanujan pour SL_2.

4. Equidistribution des points CM des variétés de Shimura

Nous devons préciser un peu les définitions relatives aux variétés de Shimura afin d'expliquer ce que l'on entend par l'équidistribution des points CM.

Soit (G, X) une donnée de Shimura; G est un groupe algébrique réductif sur $\mathbb{Q}$ et X est une $G(\mathbb{R})$ classe de conjuguaison de morphismes

$$h : \mathbb{S} \longrightarrow G_{\mathbb{R}}$$

($\mathbb{S} = \text{Res}_{\mathbb{C}/\mathbb{R}}\mathbb{G}_m$ est le tore de Deligne) vérifiant les 3 propriétés de Deligne [10] [11]. Les composantes irréductibles de X sont alors des domaines symétriques hermitiens.

Soient $\mathbb{A}_f$ l'anneau des adèles finies de $\mathbb{Q}$ et K un sous-groupe compact ouvert de $G(\mathbb{A})$, on définit sur le corps $\mathbb{C}$ la variété de Shimura

$$Sh_K(G, X) = G(\mathbb{Q})\backslash X \times G(\mathbb{A}_f)/K.$$

On vérifie que $Sh_K(G, X)$ est une réunion finie de quotients de composantes irréductibles de X par des sous-groupes de congruences de $G(\mathbb{Q})$. Par ailleurs $Sh_K(G, X)$ a un "modèle canonique" sur un corps de nombres $E(G, X)$ ne dépendant que de la donnée de Shimura (G, X).

Soit (G_1, X_1) une sous-donnée de Shimura de (G, X), on dispose alors d'une application canonique

$$f : Sh_{K\cap G_1(\mathbb{A}_f)} \longrightarrow Sh_K(G, X).$$

Une sous-variété de type de Hodge est une composante irréductible d'un translaté de l'image d'un tel morphisme par une correspondance de Hecke. (Moonen [15] caractérise ces sous-variétés en termes de variations de structures de Hodge, d'où le nom.)

Pour $h : \mathbb{S} \to G_{\mathbb{R}}$, $h \in X$, on définit le groupe de Mumford-Tate $MT(h)$ de h comme le plus petit $\mathbb{Q}$-sous-groupe H de G tel que h se factorise par $H_{\mathbb{R}}$. Si $MT(h)$ est un tore, on dit que h est spécial. Les points spéciaux de $Sh_K(G, X)$ sont les points de la forme $[h, gK]$ avec $g \in G(\mathbb{A}_f)$ et h spécial.

Fixons $h_0 \in X$ un élément spécial et $T_0 = MT(h_0)$. L'ensemble

$$S(h_0) = \{[h_0, gK], \quad g \in G(\mathbb{A}_f)\}$$

est appelé ensemble des points spéciaux de "type h_0" de X. On a une action de $T_0(\mathbb{A}_f)$ sur $S(h_0)$ donnée par $t.[h_0, gK] = [h_0, tgK]$. Pour tout $g \in G(\mathbb{A}_f)$, l'orbite sous $T_0(\mathbb{A}_f)$ de $[h_0, gK]$ est finie, on appelle "orbite torique" de $[h_0, gK]$ cette orbite. La première question naturelle est

Question 4.1 Soit $x_n = [h_n, g_n K]$ une suite générique de points spéciaux de $S = Sh_K(G, X)$. Est-il vrai que l'orbite torique de x_n est équidistribuée pour la mesure invariante normalisée de $Sh_K(G, X)$.

Notons qu'il n'est déjà pas à priori évident de prévoir la proportion des points de l'orbite torique dans les composantes de S. Il est peut-être plus réaliste de travailler dans chaque composante connexe de S (comme dans la dernière partie de ce texte). Nous tairons dans la suite ces problèmes de non connexité.

Les premiers résultats pour ces questions sont dus à Duke [12] pour la courbe modulaire $Y(1) = SL(2, \mathbb{Z})\backslash\mathbb{H}$. Il montre l'équidistribution des points à multiplication complexe par l'anneau des entiers O_K quand le discriminant tend vers l'infini. Nous expliquons dans [4], en utilisant en plus des résultats sur l'équidistribution des points de Hecke comment obtenir l'équidistribution des points à multiplication complexe par un ordre arbitraire de O_K quand le discriminant tend vers l'infini. Nous pensons plus généralement que la question 4.1 est liée aux problèmes d'équidistribution des points de Hecke décrits précédemment.

Des résultats pour l'équidistribution des orbites toriques de points CM sont annoncés par S. Zhang [26] pour les courbes de Shimura et plus généralement des variétés de Shimura de type quaternionique via un avatar de la formule de Gross-Zagier. Pour les variétés modulaires de Hilbert des résultats de ce type sont annoncés indépendamment par P. Cohen [7] (par la méthode originale de Duke) et par S. Zhang.

Les méthodes pour prouver ces énoncés comportent trois étapes que l'on va décrire de manière imprécise pour la concision de ce rapport. Soit S une variété de Shimura, soit f une fonction non constante intervenant dans la décomposition spectrale de S, soit $x_n \in S$ une suite de points CM et E_n son orbite torique. On doit montrer que

$$\lim_{n\to\infty} \frac{1}{|E_n|} \sum_{y\in E_n} f(y) = \int_S f d\mu_0. \tag{1}$$

La fonction f est alors une forme automorphe. La première étape est de montrer une "formule de classe" reliant $\frac{1}{|E_n|}\sum_{y\in E_n} f(y)$ à la valeur de la fonction L de f, tordue par une forme automorphe que l'on définit à partir de E_n, au point critique. Ce type de formule est obtenu par Waldspurger [23] pour des algèbres de quaternions sur un corps de nombres F et revisité par Zhang [26]dans le but d'obtenir les résultats d'équidistribution.

Une fois la formule de classe établie, on dispose d'une famille de fonctions L indexée par les entiers. On définit à partir de l'équation fonctionnelle de ces fonctions une notion de "conducteur analytique" q_n. L'hypothèse de Riemann (ou de Lindelöf) prévoit une borne en $0(q_n^\epsilon)$ pour la valeur critique de la fonction L considérée. Dans tous les exemples considérés, il est remarquable que pour montrer l'équidistribution il faut améliorer la borne triviale (donnée par le principe de convexité de Phragmen-Lindelöf). Ce genre de questions a reçu une attention considérable en théorie analytique des nombres et a été résolue dans de nombreux cas. On pourra consulter la série de papiers [13] et [14] pour une présentation des principaux résultats et applications de ce cercle d'idées. Notons que la démonstration

de l'équidistribution des orbites toriques de points CM sur les variétés modulaires de Hilbert utilise les résultats spectaculaires récents [8].

Pour les applications éventuelles à des énoncés arithmétiques, il paraît important de remplacer les orbites toriques par les orbites sous Galois. De manière générale si $[h, gK]$ est un point CM d'une variété de Shimura, $T = MT(h)$ est le tore associé et $E = E(T, h)$ est le corps reflexe de la variété de Shimura associé à la donnée de Shimura (T, h), l'action de Galois (cf [10], [11])se factorise à travers l'action de $T(\mathbb{A}_f)$ via un morphisme de réciprocité (et la théorie du corps de classe).

$$r : \mathrm{Res}_{E/\mathbb{Q}}\mathbb{G}_{m,E} \longrightarrow T$$

qui induit un morphisme non surjectif en général

$$r : \mathrm{Res}_{E/\mathbb{Q}}\mathbb{G}_{m,E}(\mathbb{A}_f) \longrightarrow T(\mathbb{A}_f).$$

On s'attend néanmoins à une réponse positive à la question suivante:

Question 4.2 Soit x_n une suite générique de points CM sur une variété de Shimura S, est-il vrai que les orbites sous Galois $O(x_n)$ sont équidistribuées dans S pour la mesure invariante?

De manière encore plus optimiste, on espère (par analogie avec le cas des variétés abéliennes) que le même résultat est encore vrai pour des suites strictes de points CM. Ce serait une conséquence de la conjecture d'André-Oort et de la question précédente. Notons que nous espérons que des résultats d'équidistribution pour les points CM soient en fait une étape pour montrer la conjecture en question . (C'est au moins ce qui se passe dans le cas des variétés abéliennes).

5. Equidistribution de sous-variétés spéciales

Cette partie décrit un travail [5] en cours de préparation en commun avec L. Clozel. Soit S une composante irréductible d'une variété de Shimura. Une conséquence géométrique frappante de la conjecture d'André et Oort est la suivante: Soit Y une sous-variété de S, il existe un ensemble fini $\{S_1, \ldots, S_r\}$ de sous-variétés spéciales avec $S_i \subset Y$ pour tout i tel que toute variété spéciale Z de S contenue dans Y est en fait contenue dans un des S_i.

Supposons que S est une composante irréductible de $Sh_K(G, X)$ pour un groupe G que l'on suppose adjoint (pour simplifier). On a vu qu'une sous-variété spéciale M est associée à une sous-donnée de Shimura (G_1, X_1). Si G_1 est semi-simple et X_1 contient un point spécial x_1 tel que le tore associé $T = MT(x_1) \subset G_1$ est tel que $T_{\mathbb{R}}$ est un tore maximal compact de G, on dit que M est fortement spéciale. Par exemple les variétés modulaires de Hilbert (associées à des corps totalement réels de degré n sur $\mathbb{Q}$) sont fortement spéciales dans l'espace de module $\mathcal{A}_n$ des variétés abéliennes principalement polarisées de dimension n. On peut montrer:

Théorème 5.1 *Soit Y une sous-variété d'une variété de Shimura S. Il existe un ensemble fini $\{S_1, \ldots, S_k\}$ de sous-variétés fortement spéciales de dimension positive $S_i \subset Y$ tel que si Z est une sous-variété fortement spéciale de dimension positive avec $Z \subset Y$ alors $Z \subset S_i$ pour un certain $i \in \{1, \ldots, k\}$.*

Notons que cet énoncé ne dit rien sur les sous-variétés spéciales de dimension 0 (les points spéciaux), notons cependant le corollaire suivant:

Corollaire 5.2 *Soit Y une sous-variété stricte de $\mathcal{A}_n$, il existe au plus un nombre fini de sous-variétés modulaires de Hilbert contenu dans Y.*

Le théorème 5.1 se déduit d'un énoncé ergodique. Toute sous-variété spéciale Z de S est muni d'une manière canonique d'une mesure de probabilité μ_Z.

Théorème 5.3 *Soit S_n une suite de sous-variétés fortement spéciales Soit μ_n la mesure de probabilité associée à S_n. Il existe une sous-variété fortement spéciale Z et une sous-suite μ_{n_k} qui converge faiblement vers μ_Z. De plus Z contient S_{n_k} pour tout k assez grand.*

On obtient la preuve du théorème 5.1 en considérant une suite de sous-variétés fortement spéciales maximales S_n parmi les sous-variétés fortement spéciales contenues dans Y. En passant à une sous-suite on peut supposer que μ_n converge faiblement vers μ_Z. Comme le support de μ_Z est contenu dans Y, on en déduit que $Z \subset Y$. Par la maximalité des S_n et le fait que $S_n \subset Z$ pour tout n assez grand, on en déduit que la suite S_n est stationaire.

On peut aussi réécrire cet énoncé avec la terminologie de [21]. On dit qu'une suite S_n de sous-variétés fortement spéciales est stricte si pour toute sous-variété fortement spéciale M de S,

$$\{n \in \mathbb{N},\ S_n \subset M\}$$

est fini. On peut d'ailleurs prendre dans cette définition M spéciale car une sous-variété spéciale contenant une sous-variété fortement spéciale est automatiquement fortement spéciale. Dans ce language le théorème 5.3 admet comme corollaire immédiat:

Corollaire 5.4 *Soit S_n une suite stricte de sous-variétés fortement spéciales de S. Soit μ_n et μ les mesures de probabilités associées sur S_n et S. La suite μ_n converge faiblement vers μ.*

On peut appliquer cet énoncé à des suites de sous-variétés fortement spéciales maximales. La condition d'être stricte signifie alors de ne pas avoir de sous-suites constantes. C'est par exemple le cas pour les variétés modulaires de Hilbert dans le modules des variétés abéliennes principalement polarisées $\mathcal{A}_n$.

La preuve du théorème 5.3 repose sur des résulats de Mozes et Shah [16] qui précisent la conjecture de Raghunathan démontrée par Ratner [20]. Si $S = \Gamma\backslash G(\mathbb{R})/K_\infty$ pour un sous-groupe compact maximal K_∞ et un réseau de congruence Γ, on note $\Gamma^+ = G(\mathbb{R})^+ \cap \Gamma$ et $\tilde{S} = \Gamma^+\backslash G(\mathbb{R})^+$. Si H est un sous-groupe semi-simple de $G(\mathbb{R})^+$ tel que $\Gamma^+ \cap H$ est un réseau de H alors $M_H = \Gamma^+ \cap H\backslash H$ est fermé dans $\tilde{S}$ et est muni canoniquement d'une mesure de probabilité H-invariante μ_H.

Si M_{H_n} est une suite de telles sous-variétés de $\tilde{S}$, le théorème de Mozes Shah [16] permet sous certaines hypothèses; au besoin en passant à une sous-suite; de montrer la convergence faible de μ_{H_n} vers une mesure μ_H canoniquement associée

à un $M_H = \Gamma^+ \cap H \backslash H$. En général les sous-groupes H_n n'induisent pas de sous-variétés spéciales sur S car H_n n'est pas toujours réductif et même si H_n est réductif l'espace symétrique associé à H_n n'a aucune raison d'être hermitien. Un des points clefs de la démonstration est de vérifier que si les H_n induisent des sous-variétés fortement spéciales il en est de même pour H. Pour passer de résultats sur $\tilde{S}$ à des résultats sur S on utilise aussi des résultats de Dani et Margulis ([9]thm. 2) qui donnent des critères de retour vers des compacts pour des flots unipotents sur $\tilde{S}$.

References

[1] P. Autissier *Points entiers et Théorèmes de Bertini arithmétiques.* J. Reine Angew. Math.**531**, (2001), 201–235.

[2] Y. Bilu *Limit distribution of small points on algebraic tori.* Duke Math. J. **89** (1997), n.o **3**, 465–476.

[3] L. Clozel, H. Oh, E. Ullmo. *Hecke operators and equidistribution of Hecke points.* Invent. Math., **144**, (2001), 327–351.

[4] L. Clozel, E. Ullmo. *Equidistribution des points de Hecke.* à paraitre dans "Contributions to Automorphic Forms, Geometry and Arithmetic" volume en l'honneur de Shalika, Johns Hopkins University Press, editeurs: Hida, Ramakrishnan et Shaidi.

[5] L. Clozel, E. Ullmo. *Équidistribution de sous-variétés spéciales.* En préparation.

[6] A. Chambert-Loir *Points de petite hauteur sur les variétés semi-abéliennes.* Ann. Ecole Norm. Sup. **33**, (2000) no.6, 789–821.

[7] P. Cohen. Travail en prépararation.

[8] J. Cogdel, I.I Piateskii-Shapiro, P. Sarnak. En préparartion.

[9] S.G Dani, G.A Margulis. *Limit distribution of orbits of unipotent flows and values of quadratic forms.* Adv. Sov. Math. **16**, (1993), 91–137.

[10] P. Deligne. *Travaux de Shimura.* Séminaire Bourbaki, Exposé 389, Fevrier 1971, Lecture Notes in Maths. **244**, Springer-Verlag, Berlin 1971, 123–165.

[11] P. Deligne. *Variétés de Shimura: interprétation modulaire et techniques de construction de modèles canoniques.* dans *Automorphic Forms, Representations, and L-functions* part. **2**; Editeurs: A. Borel et W Casselman; Proc. of Symp. in Pure Math. **33**, American Mathematical Society, 1979, 247–290.

[12] W. Duke. *Hyperbolic distribution problems and half-integral weight Maass forms.* Invent. math. **92**, (1988), 73–90.

[13] W. Duke, J. Friedlander, H. Iwaniec. *Bounds for automorphic L-functions I, II, III.* Invent. Math **112** (1993) No. **1**, 1–8; Invent. Math. bf 115, No **2** (1994), 219–239; Invent. Math. **143** (2001) No.2, 221-248.

[14] J. Friedlander. *Bounds for L-functions.* Proceedings of the International Congress of Mathematicians, (Zürich 1994), Birkhäuser (1995), Basel, 363–373.

[15] B. Moonen. *Linearity properties of Shimura varieties I.* Journal of Algebraic Geometry **7** (1998), 539–567.

[16] S. Mozes, N. Shah *On the space of ergodic invariant measures of unipotent flows.* Ergod. Th. and Dynam. Sys. **15**, (1995), 149–159.

[17] H. Oh. *Uniform Pointwise bounds for matrix coefficients of Unitary representations and applications to Kasdhan constants.* To appear in Duke Math. Journal.
[18] M. Raynaud. *Courbe sur une variété abélienne et points de torsion.* Invent. Math. **71**, (1983), 207–223.
[19] M. Raynaud. *Sous-variété d'une variété abélienne et points de torsion.* Arithmetic and Geometry, Paper dedicated to I. R. Shafarevich on the ocasion of his sixties birthday, vol **1**, J. Coates, S. Helgason editors. (1983) Birkhäuser.
[20] M. Ratner. *On Raghunathan's measure conjecture*, Ann. Math. **134**, (1991), 545–607.
[21] L. Szpiro, E. Ullmo, S. Zhang *Equirépartition des petits points.* Invent. Math **127**, 337–347 (1997).
[22] E. Ullmo. *Positivité et dicrétion des points algébriques des courbes.* Ann. of Maths, **147** (1998), 167–179.
[23] J.-L. Waldspurger. *Sur les valeurs de certaines fonction L automorphes en leur centre de symétrie.* Compositio Math. **54** (1985), 173–242.
[24] S. Zhang. *Equidistribution of small points on abelian varieties.* Ann. of Maths, **147**, (1998), 159–165.
[25] S. Zhang. *Small points and Arakelov theory.* Proceedings of the International Congress of Mathematicains, Vol II (Berlin 1998); Doc. Math. (1998) Extra Vol II, 217–225.
[26] S. Zhang. *Gross-Zagier formula for* GL_2. Asian J. Math. **5**, (2001), 183–290.

ICM 2002 · Vol. II · 207–217

Diophantine Methods for Exponential Sums, and Exponential Sums for Diophantine Problems

Trevor D. Wooley*

Abstract

Recent developments in the theory and application of the Hardy-Littlewood method are discussed, concentrating on aspects associated with diagonal diophantine problems. Recent efficient differencing methods for estimating mean values of exponential sums are described first, concentrating on developments involving smooth Weyl sums. Next, arithmetic variants of classical inequalities of Bessel and Cauchy-Schwarz are discussed. Finally, some emerging connections between the circle method and arithmetic geometry are mentioned.

2000 Mathematics Subject Classification: 11P55, 11L07, 11P05, 11D72, 14G05.
Keywords and Phrases: The Hardy-Littlewood method, Exponential sums, Waring's problem, Equations in many variables, Rational points, Representation problems.

1. Introduction

Over the past fifteen years or so, the Hardy-Littlewood method has experienced a renaissance that has left virtually no facet untouched in its application to diophantine problems. Our purpose in this paper is to sketch what might be termed the past, present, and future of these developments, concentrating on aspects associated with diagonal diophantine problems, and stressing modern developments that make increasing use of less traditional diophantine input within ambient analytic methods. We avoid discussion of the Kloosterman method and its important recent variants (see [5] and [8]), because the underlying ideas seem inherently constrained to quadratic, and occasionally cubic, diophantine problems. Our account begins with a brief introduction to the Hardy-Littlewood (circle) method, using

*Department of Mathematics, University of Michigan, East Hall, 525 East University Avenue, Ann Arbor, MI 48109-1109, USA. E-mail: wooley@umich.edu

Waring's problem as the basic example. The discussion here illustrates well the issues involved in the analysis of systems of diagonal equations over arbitrary algebraic extensions of $\mathbb{Q}$, and motivates that associated with more general systems of homogeneous equations (see [1] and [14]).

Let s and k be natural numbers with $s > k \geq 2$, and consider an integer n sufficiently large in terms of s and k. The circle method employs Fourier analysis in order to obtain asymptotic information concerning the number, $R(n) = R_{s,k}(n)$, of integral solutions of the equation $x_1^k + \cdots + x_s^k = n$. Write $P = n^{1/k}$ and define the exponential sum $f(\alpha) = f(\alpha; P)$ by

$$f(\alpha) = \sum_{1 \leq x \leq P} e(\alpha x^k),$$

wherein $e(z)$ denotes $e^{2\pi iz}$. Then it follows from orthogonality that

$$R(n) = \int_0^1 f(\alpha)^s e(-n\alpha) d\alpha.$$

When α is well-approximated by rational numbers with small denominators, one has sharp asymptotic information concerning $f(\alpha)$. In order to be precise, let Q satisfy $1 \leq Q \leq \frac{1}{2}P^{k/2}$, and define the *major arcs* $\mathfrak{M} = \mathfrak{M}(Q)$ to be the union of the intervals $\mathfrak{M}(q,a) = \{\alpha \in [0,1) : |q\alpha - a| \leq QP^{-k}\}$, with $0 \leq a \leq q \leq Q$ and $(a,q) = 1$. Also, put

$$S(q,a) = \sum_{r=1}^{q} e(ar^k/q) \quad \text{and} \quad v(\beta) = \int_0^P e(\beta\gamma^k) d\gamma,$$

and define $f^*(\alpha)$ for $\alpha \in [0,1)$ by taking $f^*(\alpha) = q^{-1}S(q,a)v(\alpha - a/q)$, when α lies in $\mathfrak{M}(q,a) \subseteq \mathfrak{M}(Q)$, and otherwise by setting $f^*(\alpha) = 0$. Then the sharpest available estimate (see Theorem 4.1 of [16]) establishes that[1] $f(\alpha) = f^*(\alpha) + O(Q^{1/2+\epsilon})$, uniformly for $\alpha \in \mathfrak{M}(Q)$. The functions $S(q,a)$ and $v(\beta)$ are rather well-understood, and thus one deduces that whenever $s \geq \max\{4, k+1\}$ and $Q \leq P$, then

$$\int_{\mathfrak{M}} f(\alpha)^s e(-n\alpha) d\alpha = \frac{\Gamma(1+1/k)^s}{\Gamma(s/k)} \mathfrak{S}_{s,k}(n) n^{s/k-1} + O(n^{s/k-1-\delta}), \tag{1.1}$$

for a suitable positive number δ. Here, the Γ-function is that familiar from classical analysis, and the *singular series* $\mathfrak{S}_{s,k}(n)$ is equal to the product of p-adic densities $\prod_p v_p(n)$, where for each prime p we write

$$v_p(n) = \lim_{h \to \infty} p^{h(1-s)} \mathrm{card}\{\mathbf{x} \in (\mathbb{Z}/p^h\mathbb{Z})^s : x_1^k + \cdots + x_s^k \equiv n \pmod{p^h}\}.$$

[1] Given a complex-valued function $f(t)$ and positive function $g(t)$, we use Vinogradov's notation $f(t) \ll g(t)$, or Landau's notation $f(t) = O(g(t))$, to mean that when t is large, there is a positive number C for which $f(t) \leq Cg(t)$. Similarly, we write $f(t) \gg g(t)$ when $g(t) \ll f(t)$, and $f(t) \asymp g(t)$ when $f(t) \ll g(t) \ll f(t)$. Also, we write $f(t) = o(g(t))$ when as $t \to \infty$, one has $f(t)/g(t) \to 0$. Finally, we use the convention that whenever ϵ occurs in a formula, then it is asserted that the statement holds for each fixed positive number ϵ.

An asymptotic formula for $R(n)$, with leading term determined by the major arc contribution (1.1), now follows provided that the corresponding contribution arising from the minor arcs $\mathfrak{m} = [0,1) \setminus \mathfrak{M}$ is asymptotically smaller. Although such is conjectured to hold as soon as $s \geq \max\{4, k+1\}$, this is currently known only for larger values of s. It is here that energy is focused in current research. One typically estimates the minor arc contribution via an inequality of the type

$$\left| \int_{\mathfrak{m}} f(\alpha)^s e(-n\alpha) d\alpha \right| \leq \left(\sup_{\alpha \in \mathfrak{m}} |f(\alpha)| \right)^{s-2t} \int_0^1 |f(\alpha)|^{2t} d\alpha. \tag{1.2}$$

For suitable choices of t and Q, one now seeks bounds of the shape

$$\sup_{\alpha \in \mathfrak{m}} |f(\alpha)| \ll P^{1-\tau+\epsilon} \quad \text{and} \quad \int_0^1 |f(\alpha)|^{2t} d\alpha \ll P^{2t-k+\delta+\epsilon}, \tag{1.3}$$

with $\tau > 0$ and δ small enough that $(s-2t)\tau > \delta$. The right hand side of (1.2) is then $o(n^{s/k-1})$, which is smaller than the main term of (1.1) whenever $\mathfrak{S}_{s,k}(n) \gg 1$. The latter is assured provided that non-singular p-adic solutions can be found for each prime p, and in any case when $s \geq 4k$. Classically, one has two apparently incompatible approaches toward establishing the estimates (1.3). On one side is the differencing approach introduced by Weyl [23], and pursued by Hua [9], that yields an asymptotic formula for $R(n)$ whenever $s \geq 2^k + 1$. The ideas introduced by Vinogradov [21], meanwhile, provide the desired asymptotic formula when $s > Ck^2 \log k$, for a suitable positive constant C.

2. Efficient differencing and smooth Weyl sums

Since the seminal work of Vaughan [15], progress on diagonal diophantine problems has been based, almost exclusively, on the use of smooth numbers, by which we mean integers free of large prime factors. In brief, one seeks serviceable substitutes for the estimates (1.3) with the underlying summands restricted to be smooth, the hope being that this restriction might lead to sharper bounds. Before describing the kind of conclusions now available, we must introduce some notation. Let $\mathcal{A}(P,R)$ denote the set of natural numbers not exceeding P, all of whose prime divisors are at most R, and define the associated exponential sum $h(\alpha) = h(\alpha; P, R)$ by

$$h(\alpha; P, R) = \sum_{x \in \mathcal{A}(P,R)} e(\alpha x^k).$$

When t is a positive integer, we consider the mean value $S_t(P,R) = \int_0^1 |h(\alpha)|^{2t} d\alpha$, which, by orthogonality, is equal to the number of solutions of the diophantine equation $x_1^k + \cdots + x_t^k = y_1^k + \cdots + y_t^k$, with $x_i, y_i \in \mathcal{A}(P,R)$ $(1 \leq i \leq t)$. We take $R \asymp P^\eta$ in the ensuing discussion, with η a small positive number[2]. In these

[2]We adopt the convention that whenever η appears in a statement, implicitly or explicitly, then it is asserted that the statement holds whenever $\eta > 0$ is sufficiently small in terms of ϵ.

circumstances one has $\mathrm{card}(\mathcal{A}(P,R)) \sim c(\eta)P$, where the positive number $c(\eta)$ is given by the Dickman function, and it follows that $S_t(P,R) \gg P^t + P^{2t-k}$. It is conjectured that in fact $S_t(P,R) \ll P^{\epsilon}(P^t + P^{2t-k})$. We refer to the exponent λ_t as being *permissible* when, for each $\epsilon > 0$, there exists a positive number $\eta = \eta(t,k,\epsilon)$ with the property that whenever $R \le P^{\eta}$, one has $S_t(P,R) \ll P^{\lambda_t+\epsilon}$. One expects that the exponent $\lambda_t = \max\{t, 2t-k\}$ should be permissible, and with this in mind we say that δ_t is an *associated exponent* when $\lambda_t = t + \delta_t$ is permissible, and that Δ_t is an *admissible exponent* when $\lambda_t = 2t - k + \Delta_t$ is permissible.

The computations required to determine sharp permissible exponents for a specific value of k are substantial (see [20]), but for larger k one may summarise some general features of these exponents. First, for $0 \le t \le 2$ and $k \ge 2$, it is essentially classical that the exponent $\delta_t = 0$ is associated, and recent work of Heath-Brown [6] provides the same conclusion also when $t = 3$ and $k \ge 238,607,918$. When $t = o(\sqrt{k})$, one finds that associated exponents exhibit *quasi-diagonal behaviour*, and satisfy the property that $\delta_t \to 0$ as $k \to \infty$. To be precise, Theorem 1.3 of [28] shows that whenever $k \ge 3$ and $2 < t \le 2e^{-1}k^{1/2}$, then the exponent

$$\delta_t = \frac{4k^{1/2}}{et} \exp\left(-\frac{4k}{e^2t^2}\right), \tag{2.1}$$

is associated. For larger t, methods based on repeated efficient differencing yield the sharpest estimates. Thus, the corollary to Theorem 2.1 of [26] establishes that for $k \ge 4$, an admissible exponent Δ_t is given by the positive solution of the equation $\Delta_t e^{\Delta_t/k} = ke^{1-2t/k}$. The exponent $\lambda_t = 2t - k + ke^{1-2t/k}$ is therefore always permissible. Previous to repeated efficient differencing, analogues of these permissible exponents had a term of size $ke^{-t/k}$ in place of $ke^{1-2t/k}$ (see [15]), so that in a sense, the modern theory is twice as powerful as that available hitherto.

The above discussion provides a useable analogue of the mean-value estimate in (1.3). We turn next to localised minor arc estimates. Take $Q = P$, and define $\mathfrak{m}$ as in the introduction. Suppose that s, t and w are parameters with $2s \ge k+1$ for which Δ_s, Δ_t and Δ_w are admissible exponents, and define

$$\sigma(k) = \frac{k - \Delta_t - \Delta_s\Delta_w}{2(s(k + \Delta_w - \Delta_t) + tw(1 + \Delta_s))}.$$

Then Corollary 1 to Theorem 4.2 of [27] shows that $\sup_{\alpha\in\mathfrak{m}} |h(\alpha)| \ll P^{1-\sigma(k)+\epsilon}$, and for large k this estimate holds with $\sigma(k)^{-1} = k(\log k + O(\log\log k))$. Applying an analogue of (1.2) with h in place of f, and taking[3] $t = [\frac{1}{2}k(\log k + \log\log k + 1)]$ and $s = 2t + k + [Ak\log\log k/\log k]$, for a suitable $A > 0$, we deduce from our discussion of permissible exponents that $\int_{\mathfrak{m}} h(\alpha)^s e(-n\alpha)d\alpha = o(n^{s/k-1})$. By considering the representations of a given integer n with all of the kth powers R-smooth, it is now

[3] We write $[z]$ to denote $\max\{n \in \mathbb{Z} : n \le z\}$.

apparent that a modification of the argument sketched in the introduction shows that $R(n) \gg \mathfrak{S}_{s,k}(n)n^{s/k-1}$ as soon as one confirms that

$$\int_{\mathfrak{M}} h(\alpha)^s e(-n\alpha)d\alpha \sim c(\eta)^s \frac{\Gamma(1+1/k)^s}{\Gamma(s/k)} \mathfrak{S}_{s,k}(n)n^{s/k-1}. \tag{2.2}$$

Sharp asymptotic information concerning $h(\alpha)$ is available throughout $\mathfrak{M}(Q)$ only when Q is a small power of $\log P$, and so the proof of (2.2) involves pruning technology. Such machinery, in this case designed to estimate the contribution from a set of the shape $\mathfrak{M}(P) \setminus \mathfrak{M}((\log P)^\delta)$, has evolved into a powerful tool. Such issues can be handled these days with a number of variables barely exceeding $\max\{4, k+1\}$.

This approach leads to the best known upper bounds on the function $G(k)$ in Waring's problem, defined to be the least integer r for which all sufficiently large natural numbers are the sum of at most r positive integral kth powers.

Theorem 2.1. *One has* $G(k) \leq k(\log k + \log\log k + 2 + O(\log\log k/\log k))$.

This upper bound (Theorem 1.4 of [27]) refines an earlier one of asymptotically similar strength (Corollary 1.2.1 of [24]) that gave the first sizeable improvement of Vinogradov's celebrated bound $G(k) \leq (2+o(1))k\log k$, dating from 1959 (see [22]). Aside from Linnik's bound $G(3) \leq 7$ (see [11]), all of the sharpest known bounds on $G(k)$ for smaller k are established using variants of these methods. Thus one has $G^{\#}(4) \leq 12$ (see [15], and here the # denotes that there are congruence conditions modulo 16), $G(5) \leq 17$, $G(6) \leq 24$, $G(7) \leq 33$, $G(8) \leq 42$, $G(9) \leq 50$, $G(10) \leq 59$, $G(11) \leq 67$, $G(12) \leq 76$, $G(13) \leq 84$, $G(14) \leq 92$, $G(15) \leq 100$, $G(16) \leq 109$, $G(17) \leq 117$, $G(18) \leq 125$, $G(19) \leq 134$, $G(20) \leq 142$ (see [17], [18], [19], [20]).

Unfortunately, shortage of space obstructs any but the crudest account of the ideas underlying the proof of the mean value estimates that supply the above permissible exponents. The use of exponential sums over smooth numbers occurs already in work of Linnik and Karatsuba (see [10]), but only with Vaughan's new iterative method [15] is a flexible homogeneous approach established. An alternative formulation suitable for repeated efficient differencing is introduced by the author in [24]. Suppose that the exponent λ_s is permissible, and consider a polynomial $\psi \in \mathbb{Z}[t]$ of degree $d \geq 2$. Given positive numbers M and T with $M \leq T$, and an element $x \in \mathcal{A}(T,R)$ with $x > M$, there exists an integer m with $m \in [M, MR]$ for which $m|x$. Consequently, by applying a *fundamental lemma* of combinatorial flavour, one may bound the number of integral solutions of the equation

$$\psi(z) - \psi(w) = \sum_{i=1}^{s}(x_i^k - y_i^k), \tag{2.3}$$

with $1 \leq z, w \leq P$ and $x_i, y_i \in \mathcal{A}(T,R)$ $(1 \leq i \leq s)$, in terms of the number of integral solutions of the equation

$$\psi(z) - \psi(w) = m^k \sum_{i=1}^{s}(u_i^k - v_i^k), \tag{2.4}$$

with $1 \le z, w \le P$, $M < m \le MR$, $(\psi'(z)\psi'(w), m) = 1$ and $u_i, v_i \in \mathcal{A}(T/M, R)$ $(1 \le i \le s)$. The implicit congruence condition $\psi(z) \equiv \psi(w) \pmod{m^k}$ may be analytically refined to the stronger one $z \equiv w \pmod{m^k}$, and in this way one is led to replace the expression $\psi(z) - \psi(w)$ by the difference polynomial $\psi_1(z; h; m) = m^{-k}(\psi(z + hm^k) - \psi(z))$. Notice that when $M \ge P^{1/k}$, one is forced to conclude that $z = w$, and then the number of solutions of (2.4) is bounded above by $PMRS_s(T/M, R) \ll P^{1+\epsilon}M(T/M)^{\lambda_s}$. Otherwise, following an application of Schwarz's inequality to the associated mean value of exponential sums, one may recover an equation of the shape (2.3) in which $\psi(z)$ is replaced by $\psi_1(z)$, and T is replaced by T/M, and repeat the process once again. This gives a repeated differencing process that hybridises that of Weyl with the ideas of Vinogradov.

It is now possible to describe a strategy for bounding a permissible exponent λ_{s+1} in terms of a known permissible exponent λ_s. We initially take $T = P$ and $\psi(z) = z^k$, and observe that $S_{s+1}(P, R)$ is bounded above by the number of solutions of (2.3). We apply the above efficient differencing process successively with appropriate choices for M at each stage, say $M = P^{\phi_i}$, with $0 \le \phi_i \le 1/k$, for the ith differencing operation. After some number of steps, say j, we take $\phi_j = 1/k$ in order to force the above diagonal situation that is easily estimated. One then optimises choices for the ϕ_i in order to extract the sharpest upper bound for $S_{s+1}(P, R)$, and this in turn yields a permissible exponent λ_{s+1}. It transpires that in this simplified treatment, successive admissible exponents are related by the formula $\Delta_{s+1} = \Delta_s(1 - \phi) + k\phi - 1$, wherein one may take ϕ very close to $1/(k + \Delta_s)$. Thus one finds that Δ_{s+1} is essentially $\Delta_s(1 - 2/(k + \Delta_s))$, an observation that goes some way to explaining how it is that this method is about twice as strong as previous approaches that would correspond to choices of ϕ close to $1/k$.

Refined versions of this differencing process make use of all known permissible exponents λ_s in order to estimate a particular exponent λ_t, and in such circumstances the process becomes highly iterative, and entails significant computation. Such variants make use of refined Weyl estimates for difference polynomials, and estimates for the number of integral points on curves and surfaces (see [20]). Variants of these methods apply also in the situation of Vinogradov's mean value theorem (see [25]), smooth Weyl sums with polynomial arguments (see [29]), and even for sums relevant to counting rational lines on hypersurfaces (see [12]).

Frequent reference to underlying diophantine equations seems to limit these methods to estimating even moments of smooth Weyl sums, and until recently fractional moments could be estimated only by applying Hölder's inequality to interpolate linearly between permissible exponents. However, a method [28] is now available that permits fractional moments to be estimated non-trivially, thereby "breaking classical convexity", and moreover the number of variables being differenced need not even be an integer. These new estimates can be applied to sharpen permissible exponents (with integral argument), and indeed the associated exponent (2.1) is established in this way. Another consequence [32] of these developments is the best available lower bound for $N(X)$, which we define to be the number of

integers not exceeding X that are represented as the sum of three positive integral cubes. One has $N(X) \gg X^{1-\xi/3-\epsilon}$, where $\xi = (\sqrt{2833} - 43)/41 = 0.24941301\ldots$ arises from the permissible exponent $\lambda_3 = 3 + \xi$ for $k = 3$. Earlier, Vaughan [15] obtained an estimate of the latter type with $13/4$ in place of $3 + \xi$.

3. Arithmetic variants of Bessel's inequality

Already in our opening paragraph we alluded to some of the applications accessible to the methods of §2. We now turn to less obvious applications that have experienced recent progress. We illustrate ideas once again with a simple example, and consider the set $\mathcal{Z}(N)$ of integers n, with $N/2 < n \le N$, that are *not* represented as the sum of s positive integral kth powers. The standard approach to estimating $Z(N) = \mathrm{card}(\mathcal{Z}(N))$ is via Bessel's inequality. We now take $P = N^{1/k}$. When $\mathfrak{B} \subseteq [0,1)$, write $R^*(n;\mathfrak{B}) = \int_{\mathfrak{B}} h(\alpha)^s e(-n\alpha)d\alpha$, and write also $R^*(n) = R^*(n;[0,1))$. The theory of §2 ensures that when Q is a sufficiently small power of $\log P$, and $s \ge 4k$, then $R^*(n;\mathfrak{M}) \asymp n^{s/k-1}$. Under such circumstances, an application of Bessel's inequality reveals that $Z(N)$ is bounded above by

$$\sum_{N/2<n\le N} \left| \frac{R^*(n) - R^*(n;\mathfrak{M})}{R^*(n;\mathfrak{M})} \right|^2 \ll (N^{s/k-1})^{-2} \sum_{n\in\mathbb{N}} \left| \int_{\mathfrak{m}} h(\alpha)^s e(-n\alpha)d\alpha \right|^2$$
$$\ll (N^{s/k-1})^{-2} \int_{\mathfrak{m}} |h(\alpha)|^{2s} d\alpha. \tag{3.1}$$

When $s \ge \frac{1}{2}k(\log k + \log\log k + 2 + o(1))$, the minor arc integral in (3.1) is $o(N^{2s/k-1})$, and thus it follows that $Z(N) = o(N)$. Thus one may conclude that almost all integers are sums of $s \sim (\frac{1}{2} + o(1))k\log k$ positive integral kth powers.

The application of Bessel's inequality in (3.1) makes inefficient use of underlying arithmetic information, and fails, for example, to effectively estimate the number of values of a polynomial sequence not represented in some prescribed form. Suppose instead that we define a Fourier series over the exceptional set itself, namely $K(\alpha) = \sum_n e(n\alpha)$, where the summation is over $n \in \mathcal{Z}(N)$. Since $R^*(n) = 0$ for $n \in \mathcal{Z}(N)$, one has $R^*(n;\mathfrak{m}) = -R^*(n;\mathfrak{M})$, and thus we see that

$$N^{s/k-1}Z(N) \ll \int_{\mathfrak{M}} h(\alpha)^s K(-\alpha)d\alpha = \left| \int_{\mathfrak{m}} h(\alpha)^s K(-\alpha)d\alpha \right|.$$

Applying Schwarz's inequality in combination with Parseval's identity, we recover the previous consequence of Bessel's inequality via the bound

$$\left| \int_{\mathfrak{m}} h(\alpha)^s K(-\alpha)d\alpha \right| \le \left(\int_0^1 |K(\alpha)|^2 d\alpha \right)^{1/2} \left(\int_{\mathfrak{m}} |h(\alpha)|^{2s} d\alpha \right)^{1/2}. \tag{3.2}$$

However, this formulation permits alternate applications of Schwarz's inequality or Hölder's inequality. For example, the left hand side of (3.2) is bounded above by

$$\left(\int_0^1 |h(\alpha)^{2t} K(\alpha)^2| d\alpha \right)^{1/2} \left(\int_{\mathfrak{m}} |h(\alpha)|^{2s-2t} d\alpha \right)^{1/2}, \tag{3.3}$$

and also by

$$\left(\int_0^1 |K(\alpha)|^4 d\alpha\right)^{1/4} \left(\int_{\mathfrak{m}} |h(\alpha)|^{4s/3} d\alpha\right)^{3/4}. \tag{3.4}$$

In either case, the diophantine equations underlying the integrals on the left hand sides of (3.3) and (3.4) contain arithmetic information that can be effectively exploited whenever the set $\mathcal{Z}(N)$ is reasonably thin.

The strategy sketched above has been exploited by Brüdern, Kawada and Wooley in a series of papers devoted to additive representation of polynomial sequences. Typical of the kind of results now available is the conclusion [3] that almost all values of a given integral cubic polynomial are the sum of six positive integral cubes. Also, Wooley [30], [31], has derived improved (slim) exceptional set estimates in Waring's problem when excess variables are available. For example, write $E(N)$ for the number of integers n, with $1 \le n \le N$, for which the anticipated asymptotic formula *fails* to hold for the number of representations of an integer as the sum of a square and five cubes of natural numbers. Then in [31] it is shown that $E(N) \ll N^{\epsilon}$.

As a final illustration of such ideas, we highlight an application to the solubility of pairs of diagonal cubic equations. Fix $k = 3$, define $h(\alpha)$ as in §2, and put $c(n) = \int_0^1 |h(\alpha)|^5 e(-n\alpha) d\alpha$ for each $n \in \mathbb{N}$. Brüdern and Wooley [4] have applied the ideas sketched above to estimate the frequency with which large values of $|c(n)|$ occur, and thereby have shown that, with ξ defined as in the previous section,

$$\sum_{x,y \in \mathcal{A}(P,R)} |c(x^3 - y^3)|^2 = \int_0^1 \int_0^1 |h(\alpha)^5 h(\beta)^5 h(\alpha+\beta)^2| d\alpha d\beta \ll P^{6+\xi+\epsilon}.$$

On noting that $6 + \xi < 6.25$, cognoscenti will recognise that this twelfth moment of smooth Weyl sums, in combination with a classical exponential sum equipped with Weyl's inequality, permits the discussion of pairs of diagonal cubic equations in 13 variables via the circle method. The exponent $6 + \xi$ improves an exponent $6 + 2\xi$ previously available for a (different) twelfth moment. Brüdern and Wooley [4] establish the following conclusion.

Theorem 3.1. *Suppose that $s \ge 13$, and that a_i, b_i $(1 \le i \le s)$ are fixed integers. Then the Hasse principle holds for the pair of equations*

$$a_1 x_1^3 + \cdots + a_s x_s^3 = b_1 x_1^3 + \cdots + b_s x_s^3 = 0.$$

The condition $s \ge 13$ improves on the previous bound $s \ge 14$ due to Brüdern [2], and achieves the theoretical limit of the circle method for this problem.

4. Arithmetic geometry via descent

Let $F(\mathbf{x}) \in \mathbb{Z}[x_1, \ldots, x_s]$ be a homogeneous polynomial of degree d, and consider the number, $N(B)$, of integral zeros of the equation $F(\mathbf{x}) = 0$, with

$\mathbf{x} \in [-B, B]^s$. When s is sufficiently large in terms of d, the circle method shows under modest geometric conditions that $N(B)$ is asymptotic to the expected product of local densities. For fairly general polynomials, the condition on s is as severe as $s > (d-1)2^d$, though for diagonal equations the methods of §2 relax this condition to $s > (1 + o(1))d \log d$. However, there is a class of varieties with small dimension relative to degree, for which the circle method supplies non-trivial information concerning the density of rational points. The idea is to apply a descent process in order to interpret points on the original variety in terms of corresponding points on a new variety, with higher dimension relative to degree, more amenable to the circle method.

To illustrate this principle, consider a field extension K of $\mathbb{Q}$ of degree n with associated norm form $N(\mathbf{x}) \in \mathbb{Q}[x_1, \dots, x_n]$. Also, let l and k be natural numbers with $(k, l) = 1$, and let α be a non-zero rational number. Then Heath-Brown and Skorobogatov [7] descend from the variety $t^l(1-t)^k = \alpha N(\mathbf{x})$ to the associated variety $aN(\mathbf{u}) + bN(\mathbf{v}) = z^n$, for suitable integers a and b. The circle method establishes weak approximation for the latter variety, and thereby it is shown that the Brauer-Manin obstruction is the only possible obstruction to the Hasse principle and weak approximation on any smooth projective model of the former variety. One can artificially construct further examples amenable to the circle method. For example, if we take linearly independent linear forms $L_i(\mathbf{x}) \in \mathbb{Q}[x_1, \dots, x_n]$ $(1 \le i \le n + r)$, then one can establish non-trivial lower bounds for the density of rational points on the variety $z^k = L_1(\mathbf{x}) \dots L_{n+r}(\mathbf{x})$ by descending to a variety that resembles a system of r diagonal forms of degree k, with constrained varying coefficients. The investigation of such matters will likely provide an active area of research into the future. In this context we point to work of Peyre [13], which addresses the interaction between descent and the circle method in some generality.

References

[1] B. J. Birch, Forms in many variables, *Proc. Roy. Soc. Ser. A* 265 (1962), 245–263.

[2] J. Brüdern, On pairs of diagonal cubic forms, *Proc. London Math. Soc.* (3) 61 (1990), 273–343.

[3] J. Brüdern, K. Kawada and T. D. Wooley, Additive representation in thin sequences, I: Waring's problem for cubes, *Ann. Sci. École Norm. Sup.* (4) 34 (2001), 471–501.

[4] J. Brüdern and T. D. Wooley, The Hasse principle for pairs of diagonal cubic equations (to appear).

[5] D. R. Heath-Brown, A new form of the circle method, and its application to quadratic forms, *J. Reine Angew. Math.* 481 (1996), 149–206.

[6] D. R. Heath-Brown, Equal sums of three powers (to appear).

[7] D. R. Heath-Brown & A. N. Skorobogatov, Rational solutions of certain equa-

tions involving norms, *Imperial College preprint* (June 2001).
[8] C. Hooley, On nonary cubic forms, *J. Reine Angew. Math.* 386 (1988), 32–98.
[9] L.-K. Hua, On Waring's problem, *Quart. J. Math. Oxford* 9 (1938), 199–202.
[10] A. A. Karatsuba, Some arithmetical problems with numbers having small prime divisors, *Acta Arith.* 27 (1975), 489–492.
[11] Ju. V. Linnik, On the representation of large numbers as sums of seven cubes, *Mat. Sb.* 12 (1943), 218–224.
[12] S. T. Parsell, Multiple exponential sums over smooth numbers, *J. Reine Angew. Math.* 532 (2001), 47–104.
[13] E. Peyre, Torseurs universels et méthode du cercle, *Rational points on algebraic varieties, Progr. Math. 199*, Birkhäuser, 2001, 221–274.
[14] W. M. Schmidt, The density of integer points on homogeneous varieties, *Acta Math.* 154 (1985), 243–296.
[15] R. C. Vaughan, A new iterative method in Waring's problem, *Acta Math.* 162 (1989), 1–71.
[16] R. C. Vaughan, *The Hardy-Littlewood Method*, Cambridge University Press, 1997.
[17] R. C. Vaughan & T. D. Wooley, Further improvements in Waring's problem, III: eighth powers, *Philos. Trans. Roy. Soc. London Ser. A* 345 (1993), 385–396.
[18] R. C. Vaughan & T. D. Wooley, Further improvements in Waring's problem, II: sixth powers, *Duke Math. J.* 76 (1994), 683–710.
[19] R. C. Vaughan & T. D. Wooley, Further improvements in Waring's problem, *Acta Math.* 174 (1995), 147–240.
[20] R. C. Vaughan & T. D. Wooley, Further improvements in Waring's problem, IV: higher powers, *Acta Arith.* 94 (2000), 203–285.
[21] I. M. Vinogradov, The method of trigonometric sums in the theory of numbers, *Trav. Inst. Math. Stekloff* 23 (1947), 109.
[22] I. M. Vinogradov, On an upper bound for $G(n)$, *Izv. Akad. Nauk SSSR Ser. Mat.* 23 (1959), 637–642.
[23] H. Weyl, Über die Gleichverteilung von Zahlen mod Eins, *Math. Ann.* 77 (1916), 313–352.
[24] T. D. Wooley, Large improvements in Waring's problem, *Ann. of Math.* (2) 135 (1992), 131–164.
[25] T. D. Wooley, On Vinogradov's mean value theorem, *Mathematika* 39 (1992), 379–399.
[26] T. D. Wooley, The application of a new mean value theorem to the fractional parts of polynomials, *Acta Arith.* 65 (1993), 163–179.
[27] T. D. Wooley, New estimates for smooth Weyl sums, *J. London Math. Soc.* (2) 51 (1995), 1–13.
[28] T. D. Wooley, Breaking classical convexity in Waring's problem: sums of cubes and quasi-diagonal behaviour, *Invent. Math.* 122 (1995), 421–451.

[29] T. D. Wooley, On exponential sums over smooth numbers, *J. Reine Angew. Math.* 488 (1997), 79–140.
[30] T. D. Wooley, Slim exceptional sets for sums of cubes, *Canad. J. Math.* 54 (2002), 417–448.
[31] T. D. Wooley, Slim exceptional sets in Waring's problem: one square and five cubes, *Quart. J. Math.* 53 (2002), 111–118.
[32] T. D. Wooley, Sums of three cubes, *Mathematika* (to appear).

Section 4. Differential Geometry

B. Andrews: *Positively Curved Surfaces in the Three-sphere* 221

Robert Bartnik: *Mass and 3-metrics of Non-negative Scalar Curvature* 231

P. Biran: *Geometry of Symplectic Intersections* 241

Hubert L. Bray: *Black Holes and the Penrose Inequality in General Relativity* 257

Xiuxiong Chen: *Recent Progress in Kähler Geometry* 273

Weiyue Ding: *On the Schrödinger Flows* 283

P. Li: *Differential Geometry via Harmonic Functions* 293

Yiming Long: *Index Iteration Theory for Symplectic Paths with Applications to Nonlinear Hamiltonian Systems* 303

Anton Petrunin: *Some Applications of Collapsing with Bounded Curvature* 315

Xiaochun Rong: *Collapsed Riemannian Manifolds with Bounded Sectional Curvature* 323

Richard Evan Schwartz: *Complex Hyperbolic Triangle Groups* 339

Paul Seidel: *Fukaya Categories and Deformations* 351

Weiping Zhang: *Heat Kernels and the Index Theorems on Even and Odd Dimensional Manifolds* 361

ICM 2002 · Vol. II · 221–230

Positively Curved Surfaces in the Three-sphere

B. Andrews*

Abstract

In this talk I will discuss an example of the use of fully nonlinear parabolic flows to prove geometric results. I will emphasise the fact that there is a wide variety of geometric parabolic equations to choose from, and to get the best results it can be very important to choose the best flow. I will illustrate this in the setting of surfaces in a three-dimensional sphere.

There are quite a few relevant results for surfaces in the sphere satisfying various kinds of curvature equations, including totally umbillic surfaces, minimal surfaces and constant mean curvature surfaces, and intrinsically flat surfaces. Parabolic flows can strengthen such results by allowing classes of surfaces satisfying curvature inequalities rather than equalities: This was first done by Huisken, who used mean curvature flow to deform certain classes of surfaces to totally umbillic surfaces. This motivates the question "What is the optimal result of this kind?" — that is, what is the weakest pointwise curvature condition which defines a class of surfaces which retracts to the space of great spheres?

The answer to this question can be guessed in view of the examples. To prove it requires a surprising choice of evolution equation, forced by the requirement that the pointwise curvature condition be preserved.

I will conclude by mentioning some other geometric situations in which strong results can be proved by choosing the best possible evolution equation.

2000 Mathematics Subject Classification: 53C44, 53C40.
Keywords and Phrases: Surfaces, Curvature, Parabolic equations.

1. Introduction

My aim in this talk is to demonstrate the use of fully nonlinear parabolic evolution equations as tools for proving results in differential geometry. I will emphasise the fact that there is a wide variety of flows which are geometrically defined and

*Centre for Mathematics and its Applications, Australian National University, ACT 0200, Australia. E-mail: andrews@maths.anu.edu.au

potentially applicable to geometric problems, and that there is great benefit to be had by choosing the flow carefully. I will focus on a particular application, relating to surfaces in the 3-sphere, but the method has much wider applicability.

There are some well-known examples of geometric evolution equations of the kind I want to consider: Eells and Sampson [8] used a heat flow to prove existence of harmonic maps into non-positively curved targets; Hamilton considered the flow of Riemannian metrics in the direction of their Ricci tensor, and proved that it deforms metrics of positive Ricci curvature on three-manifolds [12] and metrics of positive curvature operator on four-manifolds [13] to constant curvature metrics. The Ricci flow also gives results in higher dimensions, proved by Huisken [14], Nishikawa [24] and Margerin [19]–[21], if the curvature tensor is suitably pinched. The mean curvature flow of submanifolds of Euclidean space is also well-known as the gradient descent flow of the area functional, and because it arises in models of interfaces such as in annealing metals. The examples I will concentrate on are closest to the last example, as they are evolution equations describing submanifolds moving with curvature-dependent velocity. There are many parabolic flows of this kind, particularly for the codimension one (hypersurface) case: William Firey [11] introduced the motion by Gauss curvature as a model for pebbles wearing away as they tumble, and other flows which have been considered include motion by powers of Gauss curvature [28], [6], the square root of the scalar curvature [7], the harmonic mean of the principal curvatures [2]–[3], and the reciprocal of the mean curvature [17]. More generally, one can take the velocity to be a function of the principal curvatures which is monotone increasing in each argument.

This gives a huge variety of flows to choose from, so it makes sense to choose the flow carefully to suit the problem. I will illustrate a strategy for choosing the flow by asking that some desired curvature inequality be preserved under the flow.

I will begin, in the next two sections, by discussing some old results concerning surfaces in the three-sphere. This motivates the results of the later sections.

2. Constant mean curvature surfaces

There is a well-known result of Simons [27] which says that a minimal hypersurface in a S^{n+1} with the squared norm of the second fundamental form $|A|^2$ less than n is in fact totally geodesic (hence a great n-sphere). This result comes from an application of Simons' identity which relates the second derivatives of mean curvature to the Laplacian of the second fundamental form:

$$\nabla_i \nabla_j H = \Delta h_{ij} + |A|^2 h_{ij} - H h_i^p h_{pj} + H g_{ij} - n h_{ij}.$$

From this we can deduce if the hypersurface is minimal (so $H = 0$)

$$0 = \Delta |A|^2 - 2|\nabla A|^2 + 2|A|^2(|A|^2 - n).$$

If $|A|^2 < n$ at a maximum, then the maximum principle implies $|A|^2$ is identically zero, and the result follows. Also, if the maximum of $|A|^2$ is equal to n, then M must be a product $S^k(a) \times S^{n-k}(b)$ in $R^{k+1} \times R^{n+1-k}$, with radii a and b determined by the fact that M lies in $S^{n+1} \subset R^{n+2}$ and is minimal.

Simons' argument was taken up by other authors ([25], [5], [1]) in the slightly more general setting of constant mean curvature hypersurfaces. The results are similar: If the hypersurface has constant mean curvature H, and $|A|$ is bounded by a constant depending on n and H, then the hypersurface is totally umbillic, hence a geodesic sphere in S^{n+1}; if the inequality is not strict then the only extra possibilities are products of spheres. The argument is similar to that above, but complicated by the non-vanishing of the mean curvature.

Let me look closer at the situation for surfaces in the three-sphere: The intrinsic curvature of the surface is given by $1+\kappa_1\kappa_2 = 1+\frac{1}{2}H^2-\frac{1}{2}|A|^2$. If M is minimal, then $H=0$, so $|A|^2<2$ is equivalent to positivity of the intrinsic curvature. This is also true for constant mean curvature surfaces: In two dimensions, the curvature condition from [25] and [5] is equivalent to positivity of the intrinsic curvature.

3. Flat tori

The condition of positive intrinsic curvature seems natural in view of the results on constant mean curvature surfaces. For surfaces in space, positive curvature is a rather restrictive condition — a compact surface satisfying this condition is the boundary of a convex region. In the 3-sphere it seems somewhat less restrictive, as we can see by considering the 'boundary' case of flat surfaces, where there are the beautiful results of Weiner [32] and Enomoto [9] which classify flat tori in the 3-sphere by their Gauss maps. It was known for some time that there are many examples of these (see [26]), since the inverse image of any smooth curve in S^2 under the Hopf projection is a flat torus in S^3. These examples are all invariant under the action of $U(1)$ on $C^2 \simeq R^4$, but Weiner and Enomoto showed that there are many examples which are not symmetric.

The Gauss map of a surface in S^3 can be thought of in several ways: One can consider the tangent plane of the surface as a subspace of R^4, which gives a map from the surface to the Grassmannian $G_{2,4}$ of 2-planes in R^4. The latter is a metric product $S^2\times S^2$, and the projections onto each factor are called the self-dual and anti-self-dual Gauss maps. Alternatively, since S^3 is a group, one can map the unit normal of the surface by either left or right translations to the Lie algebra — this again gives two maps to S^2, and of course these are the same as before: The self-dual Gauss map is the same as the left-translation Gauss map, and the anti-self-dual Gauss map is the same as the right-translation Gauss map.

Enomoto [9] observed that if M^2 is intrinsically flat in S^3, then both Gauss maps are degenerate (their images are just curves in S^2). Weiner gave the complete classification result: The image curves γ_1 and γ_2 necessarily have zero total curvature, and if I_1 and I_2 are subintervals of γ_1 and γ_2 respectively, then $|\int_{I_1}\kappa\,ds| + |\int_{I_2}\kappa\,ds| < \pi$. Conversely, if γ_1 and γ_2 are any curves satisfying these conditions, then there is a flat torus with these curves as the images of the two Gauss maps, and the torus is unique up to motion by unit speed in the normal direction.

This gives a very large family of flat tori in the 3-sphere, and from these we see that surfaces with positive intrinsic curvature in S^3 can look quite complicated: The

surface can look metrically like a long thin cylinder with caps on the ends, placed in S^3 by 'winding around' a flat torus many times before closing off the ends.

4. Curvature flow

Curvature flow can give powerful generalisations of results like those from [27], [25] and [5]: Huisken [16] extended techniques developed earlier for convex hypersurfaces in Euclidean space [14] to prove the following result:

Theorem: *Let $M_0^n = x_0(M)$ be a hypersurface in S^{n+1} which satisfies*

$$|A|^2 < \frac{1}{n-1}H^2 + 2$$

if $n > 2$, and

$$|A|^2 < \frac{3}{4}H^2 + \frac{4}{3}$$

if $n = 2$. Then there exists a smooth family of hypersurfaces $\{M_t = x_t(M)\}_{0 \le t < T}$ which satisfy the same curvature condition and move by mean curvature flow with initial data M_0. Either $T < \infty$ and M_t is asymptotic to a family of geodesic spheres shrinking to their common centre, or $T = \infty$ and M_t approaches a great sphere.

This includes the result that there are no minimal surfaces with $|A|^2 < n$ except great spheres. It also implies the stronger statement that every hypersurface satisfying $|A|^2 < \frac{1}{n-1}H^2 + 2$ can be deformed, keeping this condition, to a great sphere (except in the case $n = 2$). The condition $|A|^2 < \frac{1}{n-1}H^2 + 2$ is the same as that arrived at by Okumura [25] for constant mean curvature surfaces (Cheng and Nakagawa [5] improved this for higher dimensions, but in two dimensions it is sharp). The proof of the above result is significantly more difficult than that for the constant mean curvature case.

The result seems very satisfying, except when $n = 2$ where the method does not seem to work for Okumura's condition $|A|^2 < H^2 + 2$. The latter is exactly the condition of positive intrinsic curvature. This raises several questions: Does mean curvature flow in fact preserve this condition? If not, is there any flow which does?

5. The optimal result

5.1. Choosing the evolution equation

Now we can illustrate the method: The previous questions can be answered in a rather systematic way. The idea is to write down the conditions required for an arbitrary flow by a function F of curvature to preserve positive intrinsic curvature.

We can write down an evolution equation for an arbitrary function G of the principal curvatures κ_1 and κ_2, and see what conditions are required for the flow to preserve the condition $G \ge 0$. For convenience we can write G in the form

$$G(\kappa_1, \kappa_2) = (\kappa_1 - \kappa_2)^2 - \varphi(\kappa_1 + \kappa_2)^2 \tag{5.1}$$

so that in the case we are interested in, $\varphi(x) = \sqrt{4 + x^2}$. We can also write

$$F = f(\kappa_1 + \kappa_2, G). \tag{5.2}$$

Then the evolution equation for G is as follows:

$$\frac{\partial G}{\partial t} = \dot{F}^{ij} \nabla_i \nabla_j G + Q(h)(\nabla h, \nabla h) + Z(h), \tag{5.3}$$

where $\dot{F}$ is the matrix of derivatives of F with respect to the components of the second fundamental form, which is positive definite as long as F is an increasing function of each of the principal curvatures. The second term is a quadratic function of the components of the derivative of the second fundamental form, with coefficients depending on curvature h, explicitly given by

$$Q = \left(\dot{G}^{ij} \ddot{F}^{kl,mn} - \dot{F}^{ij} \ddot{G}^{kl,mn} \right) \nabla_i h_{kl} \nabla_j h_{mn},$$

where $\ddot{F}$ is the second derivative of F with respect to the components of h. The last term Z depends on the curvature alone, and has the form

$$\begin{aligned} Z &= \dot{G}^{ij} \left(F(h_{ij}^2 + g_{ij}) + \dot{F}^{kl} \left(h_{ij} h_{kl}^2 - h_{kl} h_{ij}^2 + g_{ij} h_{kl} - g_{kl} h_{ij} \right) \right) \\ &= F \left(\dot{G}^1 (1 + \kappa_1^2) + \dot{G}^2 (1 + \kappa_2^2) \right) + (1 + \kappa_1 \kappa_2)(\kappa_2 - \kappa_1)(\dot{G}^1 \dot{F}^2 - \dot{F}^1 \dot{G}^2). \end{aligned}$$

To show that $G \geq 0$ is preserved (with $G = 1 + \kappa_1 \kappa_2$), we consider the situation at a point where G first attains a zero minimum. Then the first term on the right-hand side of (5.3) is non-negative; we consider each of the other terms. The last term is simplest: Substituting the forms of F and G from (5.1) and (5.2), we find

$$Z = G \left(fH + \frac{\partial f}{\partial H} \varphi^2 \right),$$

so Z vanishes at a zero of G, no matter what speed F we use. This is another indication of the fact that the condition of positive intrinsic curvature is optimal. The gradient terms are the most complicated, but we can simplify them significantly by observing two things: First, ∇h is a totally symmetric 3-tensor, by the Codazzi equation. Second, at a minimum of G, the gradients of G vanish. It follows that there are only two independent components of ∇h, and one finds that these never mix in the expression for Q, so that

$$Q = \alpha (\nabla_1 h_{22})^2 + \beta (\nabla_2 h_{11})^2.$$

Since we have no further information about ∇h (that is, no reason to expect that the magnitudes of these remaining components should vanish) we must impose the condition that α and β are non-negative. This gives two conditions, which we can interpret as conditions on the first and second derivatives of F. A fact which is perhaps not obvious is that these conditions only involve the restriction of F to the

boundary of the set $\{G = 0\}$ in the curvature plane, so we can consider F as defined by (5.2) with $G = 0$. Then the conditions can be written explicitly as follows:

$$\frac{\varphi''}{1-\varphi'} - \frac{1-\varphi'}{\varphi} \leq \frac{f''}{f'} \leq \frac{\varphi''}{1+\varphi'} + \frac{1+\varphi'}{\varphi}.$$

In the case of interest, we have $\varphi = \sqrt{4+H^2}$, and the first and last quantities are both equal to $-2H/(4+H^2)$. The only possibilities for F are the following:

$$F = C_1 + C_2 \arctan\left(\frac{H}{2}\right).$$

This applies only along the curve $\{G = 0\}$, so we are reasonably free to choose F in the region where $G > 0$, as long as it is monotone in both principal curvatures.

5.2. The extreme case

The remarkably restricted form of the evolution equation is illuminated somewhat by considering the extreme case of flat surfaces: If the flow preserves positive intrinsic curvature, then it must also preserve zero curvature. As outlined above, the structure of surfaces with zero curvature is very well understood, and in particular the Gauss map $G : M^2 \to S^2 \times S^2$ has the remarkable property that the projection onto each factor is one-dimensional. This must be preserved under the flow.

The flow we have ended up with is characterised by the fact that the Gauss map evolves according to the mean curvature flow (now for codimension 2 surfaces in $S^2 \times S^2$, which means that each of the two curves coming from the two projections of the Gauss map evolves according to the curve-shortening flow in S^2. Since each of the curves divides the area of the sphere into two equal parts, the image of the Gauss map never develops singularities (at least in the case where the two curves are homotopic to great circles traversed once), but in fact the flat tori will in general develop singularities — this is analogous to the motion of a curve in the plane with constant normal speed, which develops singularities even though the normal direction stays constant at each point. Incidentally, there has been some very impressive recent progress on mean curvature flow in higher codimension, due to Mu-Tao Wang [29]–[31], who has used it to prove several very interesting results regarding maps between manifolds.

The examples of flat tori can be used to prove that there is no other curvature-driven flow of surfaces which preserves the condition of positive curvature, by giving examples for any other flow of flat tori which do not stay flat.

5.3. Regularity

A technical issue which arises is the following: The speed we ended up with is not concave or convex as a function of the second fundamental form. The regularity estimates due to Krylov [18] and Evans [10] for fully nonlinear equations (needed to prove that we get classical solutions of the flow) require concavity, so we cannot

use these. Instead it is possible to adapt the estimates for elliptic equations in two variables (due to Morrey [22] and Nirenberg [23]) to give good $C^{2,\alpha}$ estimates for solutions of fully nonlinear parabolic equations in two space variables.

5.4. Curvature pinching

Now we come to the problem of choosing a good way to extend the speed from the boundary $\{G = 0\}$ to the interior of the region $\{G > 0\}$. The idea is to do this in such a way that any compact surface with strictly positive curvature necessarily has very strongly controlled curvature in the future — that is, we want the region $\{G > 0\}$ to be exhausted by a nested family of regions which stay away from the boundary, and only approach infinity near the 'umbillic' line $\kappa_1 = \kappa_2$. This means that any singularity which occurs will have to be totally umbillic, so occurs only when the surface shrinks to a point while becoming spherical in shape.

This can be done in many ways. One which is relatively simple to describe, but results in solutions which are only $C^{2,\alpha}$, is as follows: Take

$$F = \begin{cases} \arctan \kappa_1 + \arctan \kappa_2, & \kappa_1 \kappa_2 < 1; \\ \frac{\pi}{4}(\kappa_1 \kappa_2 + 1), & \kappa_1 \kappa_2 > 1. \end{cases}$$

This is then a Lipschitz, monotone increasing function of the curvatures, and one can check that the following regions of the curvature plane are preserved:

$$\Omega_\varepsilon = \left\{ |\kappa_1 - \kappa_2| \leq \frac{1 + \kappa_1 \kappa_2}{\varepsilon} \right\} \cap \{\kappa_1 \kappa_2 \leq 1\} \cup \left\{ |\kappa_1 - \kappa_2| \leq \frac{2}{\varepsilon} \right\} \cap \{\kappa_1 \kappa_2 \geq 1\}.$$

This means that the difference between the principal curvatures stays bounded even if the curvature becomes large, which implies very strong control on singularities. This is similar to the estimate used in [4] to prove that worn stones (i.e. convex surfaces moving by their Gauss curvature) become round as they shrink to points.

With a little more work we can choose the speed to be a smooth function of the principal curvatures, and then solutions are also smooth.

In the choice above, we also have the nice feature that minimal surfaces do not move. We can with slight modifications arrive at a speed for which constant mean curvature surfaces do not move, for any particular choice of the mean curvature, as long as we are willing to work in the category of oriented surfaces. More generally, we can contrive that for a given monotone increasing function ϕ of the principal curvatures, surfaces satisfying $\phi = 0$ do not move. Here F (and ϕ) must be symmetric. We can also choose if desired a speed which is always positive, so that there are no stationary solutions.

5.5. The results

The main result for the above speed is the following:

Theorem 1. *Let x_0 be an immersion of S^2 in S^3, with non-negative intrinsic curvature in the induced metric. Then the flow constructed above deforms $M_0 = x_0(S^2)$ through a family $M_t = x_t(S^2)$, with intrinsic curvature strictly positive for*

each $t > 0$, to either a great sphere (in infinite time) or to a point, with spherical limiting shape (in finite time). If M_0 is embedded, then so is M_t for each $t > 0$.

This includes in particular Simons' result on mimimal surfaces. If we modify the speed somewhat, then we get the following result, which gives in particular a new result for Weingarten surfaces in the 3-sphere:

Theorem 2. *Let ϕ be any smooth, strictly monotone function of κ_1 and κ_2 defined on $\{\kappa_1\kappa_2 + 1 \geq 0\}$. Then there exists a function F which is smoothly defined on $\{\kappa_1\kappa_2 + 1 \geq 0\}$, and strictly monotone increasing in each argument, with $\mathrm{sgn}F = \mathrm{sgn}\phi$ everywhere, such that the following holds: If $M_0 = x_0(S^2)$ is a smooth compact surface in S^3 with non-negative intrinsic curvature, then the motion with speed F deforms M_0 through a smooth family $\{M_t\}_{0\leq t<T}$, each strictly positively curved, which either converge to a point with spherical limiting shape with $T < 0$, or converge to a totally umbillic surface (spherical cap) with $\phi = 0$ if $T = \infty$.*

This includes two cases: Either there is some point where $\phi = 0$, in which case there is a spherical cap with $\phi = 0$ and the above result implies that this is the only surface with $\phi = 0$ with positive intrinsic curvature, or ϕ is never zero, in which case all surfaces converge to points. In the latter case a very small geodesic sphere with one choice of orientation will shrink inwards to its centre, while the same sphere with the opposite orientation expands over the equator and eventually contracts to the antipodal point. In this way we have a unique way of associating an oriented surface with the point it eventually contracts to, and we deduce the following:

Theorem 3. *The space of oriented surfaces with positive intrinsic curvature in S^3 retracts onto S^3.*

Finally, if we introduce some non-local terms in the speed, we can devise a flow which fixes the enclosed volume, preserves positive intrinsic curvature, and gives convergence to spherical caps, without moving constant mean curvature surfaces.

6. Other results by related methods

The methods I outlined above also yield interesting results for a variety of other problems: One which works out similarly, and which has some interesting parallels, is that of surfaces in three-dimensional hyperbolic space. The surfaces of interest are those for which all of the principal curvatures are less than 1 in magnitude. We can find a flow which deforms any such surface in a compact hyperbolic manifold to a minimal surface, while keeping the principal curvatures less than 1 in magnitude. Rather surprisingly, this flow is in a way the hyperbolic analogue of the one we just described for the sphere: Instead of moving with speed equal to the sum of the arctangents of the principal curvatures, we move with speed equal to the sum of the hyperbolic arctangents of the principal curvatures. The resulting flow is very well-behaved, and has the interesting property that the Gauss map of the surface (the map which takes a point of the surface to its tangent plane, thought of as a point in the Grassmannian of spacelike 2-planes in Minkowksi space $R^{3,1}$), evolves according to mean curvature flow.

The methods also give good results for hypersurfaces in higher-dimensional spheres: Hypersurfaces with positive sectional curvatures can be deformed in such a way as to preserve that condition, and similar results can be deduced. The condition of positive sectional curvature can probably be relaxed: Positive sectional curvature is implied by the condition of Okumura [25] for constant mean curvature hypersurfaces, but not by the sharper condition of Cheng and Nakagawa [5] and Alencar and do Carmo [1].

References

[1] H. Alencar and M. do Carmo, Hypersurfaces with constant mean curvature in spheres, *Proc. Amer. Math. Soc.* 120 (1994), 1223–1229.

[2] B. Andrews, Contraction of convex hypersurfaces in Euclidean space, *Calc. Var. P.D.E.* 2 (1994), 151–171.

[3] B. Andrews, Contraction of convex hypersurfaces in Riemannian spaces, *J. Differential Geometry* 39 (1994), 407–431.

[4] B. Andrews, Gauss Curvature Flow: The Fate of the Rolling Stones, *Invent. Math.* 138 (1999), 151–161.

[5] Q.-M. Cheng and H. Nakagawa, Totally umbillic hypersurfaces, *Hiroshima Math. J.* 20 (1990), 1–10.

[6] B. Chow, Deforming convex hypersurfaces by the nth root of the Gaussian curvature, *J. Differential Geom.* 22 (1985), 117–138.

[7] B. Chow, Deforming convex hypersurfaces by the square root of the scalar curvature, *Invent. Math.* 87 (1987), 63–82.

[8] J. Eells and J. Sampson, Harmonic mappings of Riemannian manifolds, *Amer. J. Math.* 86 (1964), 109–160.

[9] K. Enomoto, The Gauss image of flat surfaces in R^4, *Kodai Math. J.* 9 (1986), 19–32.

[10] L. C. Evans, Classical solutions of fully nonlinear, convex, second order elliptic equations, *Comm. Pure Appl. Math.* 24 (1982), 333–363.

[11] W. J. Firey, Shapes of worn stones. *Mathematika* 21 (1974), 1–11.

[12] R. S. Hamilton, Three-manifolds with positive Ricci curvature, *J. Differential Geometry*, 17 (1982), 255–306.

[13] Four-manifolds with positive curvature operator, *J. Differential Geometry* 24 (1986), 153–179.

[14] G. Huisken, Flow by mean curvature of convex surfaces into spheres, *J. Differential Geometry* 20 (1984), 237–266.

[15] G. Huisken, Ricci deformation of the metric on a Riemannian manifold, *J. Differential Geometry* 21 (1985), 47–62.

[16] G. Huisken, Deforming hypersurfaces of the sphere by their mean curvature, *Math. Z.* 195 (1987), 205–219.

[17] G. Huisken and T. Ilmanen, The Riemannian Penrose Inequality, *Internat. Math. Res. Notices* 1997, no. 20, 1045–1058.

[18] N. V. Krylov, Boundedly inhomogeneous elliptic and parabolic equations, *Izvestia Akad. Nauk. SSSR* 46 (1982), 487–523. English translation in *Math.*

USSR Izv. 20 (1983).

[19] C. Margerin, Pointwise pinched manifolds are space forms, *Proc. Symp. Pure Math.* 44, 1986.

[20] C. Margerin, Une caractrisation optimale de la structure diffrentielle standard de la sphre en terme de courbure pour (presque) toutes les dimensions, *C. R. Acad. Sci. Paris Sér I Math.* 319 (1994) 713–716 and 605–607.

[21] C. Margerin, A sharp characterization of the smooth 4-sphere in curvature terms, *Comm. Anal. Geom.* 6 (1998), 21–65.

[22] C.B. Morrey, Jr., On the solutions of quasi-linear elliptic partial differential equations, *Trans. Amer. Math. Soc.* 43, (1938), 126–166.

[23] L. Nirenberg, On nonlinear elliptic partial differential equations and Hölder continuity, *Comm. Pure Appl. Math.* 6 (1953), 103–156.

[24] S. Nishikawa, Deformation of Riemannian metrics and manifolds with bounded curvature ratios, *Proc. Sympos. Pure Math.* 44, 1986.

[25] M. Okumura, Hypersurfaces and a pinching problem on the second fundamental tensor, *Amer. J. Math.* 96 (1974), 207–213.

[26] U. Pinkall, Hopf Tori in S^3, *Invent. Math.* 81 (1985), 379–386.

[27] J. Simons, Minimal varieties in Riemannian manifolds, *Ann. of Math.* (2) 88 (1968), 62–105.

[28] Kaising Tso, Deforming a hypersurface by its Gauss-Kronecker curvature, *Comm. Pure Appl. Math.* 38 (1985), 867–882.

[29] M.-T. Wang, Mean curvature flow of surfaces in Einstein four-manifolds, *J. Differential Geom.* 57 (2001), 301–338.

[30] M.-T. Wang, Deforming area preserving diffeomorphism of surfaces by mean curvature flow, *Math. Res. Lett.* 8 (2001), 651–661.

[31] M.-T. Wang, Subsets of Grassmannians preserved by mean curvature flow, preprint, 2002.

[32] J. Weiner, Flat tori in S^3 and their Gauss maps, *Proc. London Math. Soc.* 62 (1991), 54–76.

ICM 2002 · Vol. II · 231–240

Mass and 3-metrics of Non-negative Scalar Curvature

Robert Bartnik*

Abstract

Physicists believe, with some justification, that there should be a correspondence between familiar properties of Newtonian gravity and properties of solutions of the Einstein equations. The Positive Mass Theorem (PMT), first proved over twenty years ago [45, 53], is a remarkable testament to this faith. However, fundamental mathematical questions concerning mass in general relativity remain, associated with the definition and properties of quasi-local mass. Central themes are the structure of metrics with non-negative scalar curvature, and the role played by minimal area 2-spheres (black holes).

2000 Mathematics Subject Classification: 53C99, 83C57.
Keywords and Phrases: Quasi-local mass, Einstein equations, Scalar curvature.

1. Positive Mass Theorem

The Positive Mass Theorem provides a good example of *"the unreasonable effectiveness of physics in mathematics*[1] *"*. The need to define mass in general relativity is motivated directly by the physics imperative to establish a correspondence between general relativity and classical Newtonian gravity. Already difficulties arise: although the vacuum Einstein equations $Ric_{\alpha\beta} - \frac{1}{2}Rg_{\alpha\beta} = 0$ for the Lorentz metric $g_{\alpha\beta}$ suggest (by analogy with the wave equation, for example) that a mass (energy) which includes contributions from the gravitational field, should be built from the first derivatives of the field $g_{\alpha\beta}$, it is clear that this is incompatible with coordinate invariance.

The Schwarzschild vacuum spacetime metric, for $r > \max(0, 2\mathrm{M})$,

$$ds^2 = -(1 - 2\mathrm{M}/r)\, dt^2 + \frac{dr^2}{1 - 2\mathrm{M}/r} + r^2(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2), \tag{1.1}$$

*School of Mathematics and Statistics, University of Canberra, ACT 2601, Australia. E-mail: robert.bartnik@canberra.edu.au

[1] with apologies to Eugene Wigner [52].

provides an important clue, since the parameter $\mathrm{M} \in \mathbb{R}$ governs the behaviour of timelike geodesics and may be regarded as the total mass. Note that $\mathrm{M} > 0$ ensures the boundary $r = 2\mathrm{M}$ is smooth and totally geodesic in the hypersurfaces $t = const.$

A Riemannian 3-manifold (M, g) is said to be *asymptotically flat* if $M \backslash K \simeq \mathbb{R}^3 \backslash B_1(0)$ for some compact K, and M admits a metric $\mathring{g}$ which is flat outside K, and the metric components g_{ij} in the induced rectangular coordinates satisfy

$$|g_{ij} - \mathring{g}_{ij}| = O(r^{-1}), \quad |\partial_k g_{ij}| = O(r^{-2}), \quad |\partial_k \partial_l g_{ij}| = O(r^{-3}). \tag{1.2}$$

The total mass of (M, g) is defined informally by [1]

$$m_{ADM} = \frac{1}{16\pi} \oint_{S^2(\infty)} (\partial_i g_{ij} - \partial_j g_{ii})\, dS_j. \tag{1.3}$$

If the scalar curvature $R(g) \in L^1(M)$ then m_{ADM} is well-defined, independent of the choices of rectangular coordinates and of exhaustion of M used to define $\oint_{S^2(\infty)}$ — see [3, 15, 37] for weaker decay and smoothness assumptions.

For simplicity, the discussion here is restricted to C^∞ Riemannian 3-dimensional geometry. This corresponds to the case of time-symmetric initial data: (M, g) is a totally geodesic spacelike hypersurface in a Lorentzian manifold, and we can identify the local matter (equivalently, energy) density with the scalar curvature $R(g) \geq 0$. This simplification entails a small loss of generality: most, but not all, of the results we describe have been extended to general asymptotically flat space-time initial data (M, g, K), where K_{ij} is the second fundamental form of a spacelike hypersurface M. Some results also generalize to the closely related Bondi mass, which measures mass and gravitational radiation flux near null infinity, and to mass on asymptotically hyperbolic and anti-deSitter spaces cf. [51, 16], but these involve additional complications which we will not discuss here.

The Positive Mass Theorem (PMT) in its simplest form is

Theorem 1 *Suppose (M, g) is a complete asymptotically flat 3-manifold with non-negative scalar curvature $R(g) \geq 0$. Then $m_{ADM} \geq 0$, and $m_{ADM} = 0$ iff $(M, g) = (\mathbb{R}^3, \delta)$.*

The rigidity conclusion in the case $m_{ADM} = 0$ shows that $m_{ADM} > 0$ for (M, g) scalar flat ("matter-free") but non-flat, so m_{ADM} does provide a measure of the gravitational field.

Three distinct approaches have been successfully used to prove the PMT: with stable minimal surfaces [45, 46]; with spinors [53, 36] and the Schrödinger-Lichnerowicz identity [48, 35]; and using the Geroch foliation condition [23, 30]. A number of other appproaches have produced partial results: using spacetime geodesics [42]; a nonlinear elliptic system for a distinguished orthonormal frame [39, 18]; and alternative foliation conditions [32, 33, 6]. The connection between these approachs remains mysterious; the only discernable common thread is mean curvature, and this is quite tenuous.

The application of the positive mass theorem to resolve the Yamabe conjecture [44, 34] is well known. Less well known is the proof of the uniqueness of the

Schwarzschild spacetime amongst static metrics with smooth black hole boundary [13], which we briefly outline.

A *static spacetime* is a Lorentzian 4-manifold with a hypersurface-orthogonal timelike Killing vector. With V denoting the length of the Killing vector, the metric g on the spacelike hypersurface satisfies the static equations

$$\begin{aligned} Ric_g &= V^{-1}\nabla^2 V, \\ \Delta_g V &= 0. \end{aligned} \tag{1.4}$$

Smoothness implies the boundary set $\Sigma = \{V = 0\}$ is totally geodesic; analyticity of g, V can be used to show the asymptotic expansions

$$\begin{aligned} g_{ij} &= (1 + 2m/r)\delta_{ij} + O(r^{-2}), \\ V &= 1 - m/r + O(r^{-2}), \end{aligned}$$

as $r \to \infty$ for some constant $m \in \mathbb{R}$. The metrics $g^{\pm} = \phi_{\pm}^4 g$ where $\phi_{\pm} = (1 \pm V)/2$ both have $R(g^{\pm}) = 0$, and g^+ is asymptotically flat with vanishing ADM mass, and g^- is a (smooth) metric on a compact manifold. Gluing two copies of (M, g) along the totally geodesic boundary Σ and conformally changing to $\tilde{g} = \tilde{\phi}^4 g$ where $\tilde{\phi} = \phi_{\pm}$ on the two copies of M, gives a complete AF manifold with $R(\tilde{g}) = 0$ and vanishing mass. The PMT shows $(\tilde{M}, \tilde{g})$ is flat and it follows without difficulty that (M, g) is Schwarzschild. This extends previous results [31, 43] which required the boundary to be connected.

2. Penrose conjecture

A boundary component Σ with mean curvature $H = 0$ is called a *black hole* or *horizon*, since if (M, g) is a totally geodesic hypersurface then Σ is a trapped surface and hence, by the Penrose singularity theorem [26], lies within an event horizon and is destined to encounter geodesic incompleteness in the predictable future.

The spatial Schwarzschild metric $g = \frac{dr^2}{1-2\mathrm{M}/r} + r^2(d\vartheta^2 + \sin\vartheta d\varphi^2)$ with $\mathrm{M} < 0$ shows that the completeness condition in the PMT is important, but it can be weakened to allow horizon boundary components of M. This follows immediately from the minimal surface argument [45]; or by an extension to the Witten argument [22], imposing one of the boundary conditions

$$\psi = \pm\epsilon\psi \text{ on } \Sigma = \partial M, \tag{2.1}$$

on the spinor field ψ, where $\epsilon = \gamma^n\gamma^0$ satisfies $\epsilon^2 = 1$. An interesting extension is obtained by imposing the spectral boundary condition

$$P_+\psi = 0 \text{ on } \Sigma \tag{2.2}$$

where P_+ is the projection onto the subspace of positive eigenspinors of the induced Dirac operator $\mathcal{D}_\Sigma$. Using the remarkable Hijazi-Bär estimate [28, 2]

$$|\lambda| \geq \sqrt{4\pi/|\Sigma|}, \tag{2.3}$$

for the eigenvalues of $\mathcal{D}_\Sigma$ when $\Sigma \simeq S^2$, Herzlich showed [27]

Theorem 2 *If (M, g) is asymptotically flat with $R(g) \geq 0$ and boundary $\Sigma \simeq S^2$ with mean curvature satisfying*

$$H_\Sigma \leq 2/r \tag{2.4}$$

where $r = \sqrt{|\Sigma|/4\pi}$, then $m_{ADM} \geq 0$, with equality iff $(M, g) = (\mathbb{R}^3 \setminus B(r), \delta)$.

The proof starts with the Riemannian form of the Schrödinger-Lichnerowicz-Witten identity [48, 35, 53]

$$\int_M (|\nabla\psi|^2 + \tfrac{1}{4}R(g)|\psi|^2 - |\mathcal{D}\psi|^2)\, dv_M = 4\pi|\psi_\infty|^2 m_{ADM} + \oint_\Sigma \mu(\psi), \tag{2.5}$$

where $\mu(\psi)$ is the Nester-Witten form [38]

$$\mu(\psi) \quad = \quad \langle \psi, (\mathcal{D}_\Sigma + \tfrac{1}{2}H_\Sigma)\psi\rangle\, dv_\Sigma. \tag{2.6}$$

The boundary condition $P^+\psi|_\Sigma = 0$ is elliptic and it can be shown [8] there is a spinor on M satisfying $\mathcal{D}\psi = 0$ with boundary conditions $\psi \to \psi_\infty \neq 0$ as $r \to \infty$ and (2.2) on Σ. It follows from (2.3) and (2.2) that $\langle \psi, (\mathcal{D}_\Sigma + \frac{1}{2}H_\Sigma)\psi\rangle \leq (\frac{1}{2}H_\Sigma - |\lambda_1^-|)|\psi|^2 \leq 0$ and the result follows.

Observe that in each case, equality leads to flat $\mathbb{R}^3$. An elegant physical argument lead Penrose to conjecture an analogous inequality, but which distinguishes the Schwarzschild metric instead [40], see also [24].

Conjecture 3 (Penrose) *If (M, g) satisfies the conditions of the PMT, except that $\partial M = \Sigma$ is compact with vanishing mean curvature and such that Σ is the "outermost" closed minimal surface in M, then*

$$m_{ADM} \geq \sqrt{|\Sigma|/16\pi}, \tag{2.7}$$

with equality only for the Schwarzschild metric.

A closed minimal surface is said to be an *outermost horizon* or *outer-minimizing horizon* if M contains no least area surfaces homologous to Σ in the asymptotic region exterior to Σ. The outermost condition is essential, since examples of non-negative scalar curvature manifolds can be constructed by forming the connected sum of M and large spheres by arbitrarily small and large necks.

The Penrose conjecture has been established by Huisken and Ilmanen [29, 30] using a variational level set formulation of the inverse mean curvature flow [23], and by Bray [12] by a very interesting conformal deformation argument. Bray's proof is more general since it takes into account contributions from all the connected components of the boundary.

3. Quasi-local mass

Thus it is natural to consider $\sqrt{|\Sigma|/16\pi}$ as the mass of a black hole (minimal surface) Σ. More generally, the correspondence with Newtonian gravity suggests that any bounded region (Ω, g) should have a *quasi-local* mass, which measures both

the matter density (represented in this case by the scalar curvature $R(g) \geq 0$), and some contribution from the gravitational field. The rather satisfactory positivity properties of the total mass, as established by the PMT, motivate the properties we might expect such a geometric mass to possess [20, 14, 7].

1. **(non-negativity)** $m_{QL}(\Omega) \geq 0$;
2. **(rigidity/strict positivity)** $m_{QL}(\Omega) = 0$ if and only if (Ω, g) is flat;
3. **(monotonicity)** $m_{QL}(\Omega_1) \leq m_{QL}(\Omega_2)$ whenever $\Omega_1 \subset \Omega_2$, where it is understood that the inclusion is a metric isometry;
4. **(spherical mass)** m_{QL} should agree with the spherical mass, for spherically symmetric regions;
5. **(ADM limit)** m_{QL} should be asymptotic to the ADM mass;
6. **(black hole limit)** m_{QL} should agree with the black hole mass (2.7).

Many candidates have been proposed for quasi-local mass (see for example [10] for a comparison of some definitions), the most significant being that of Hawking [25],

$$m_H(\Sigma) = \sqrt{\frac{|\Sigma|}{16\pi}} \left(1 - \frac{1}{16\pi} \oint_\Sigma H^2\right) \tag{3.1}$$

where $\Sigma = \partial\Omega$. This equals M for standard spheres in Schwarschild. Although $m_H \leq 0$ for surfaces in $\mathbb{R}^3$, it was shown in [14] that $m_H(\Sigma) \geq 0$ for a stable constant mean curvature 2-sphere Σ in a 3-manifold of non-negative scalar curvature. Thus for such "round" spheres, m_H is nonegative, and the black hole limit condition is trivially satisfied. However the remaining properties, in particular rigidity and monotonicity, are rather problematic. Although the twistorially-defined Penrose quasi-local mass [41] is well-behaved in special cases [50], it is defined unambiguously only for surfaces arising from embedding into a conformally flat spacetime, and even then numerical experiments [11] strongly suggest that monotonicity is violated.

In fact, of the various proposals for m_{QL}, only the definitions of [14, 5, 19] are known to satisfy positivity. Dougan and Mason [19] show that the integral $\oint_\Sigma \mu(\psi)$ of the Nester-Witten 2-form (2.6) is positive for spinor fields ψ on Σ which satisfy a certain elliptic system on Σ. However, Bergqvist [9] shows that positivity holds under much weaker conditions on ψ, and there are many variant definitions with similar properties. It would be useful to understand these DM-style definitions better, and in particular whether any satisfy monotonicity.

Monotonicity and ADM-compatibility imply $m_{QL}(\Omega) \leq m_{ADM}(M, g)$ for any region Ω embedded isometrically in an (M, g) satisfying (as always) the PMT conditions. This motivates the following definition [4, 30]

Definition 4 *Let $\mathcal{PM}$ denote the set of all asymptotically flat 3-manifolds (M, g) of non-negative scalar curvature, with boundary which if non-empty, consists of compact outermost horizons, and such that (M, g) has no other horizons. For any bounded open connected region (Ω, g), let $\mathcal{PM}(\Omega)$ be the set of $(M, g) \in \mathcal{PM}$ such that Ω embeds isometrically into M, and define*

$$m_{QL}(\Omega) = \inf\{m_{ADM}(M, g) : (M, g) \in \mathcal{PM}(\Omega)\}. \tag{3.2}$$

We say that M satisfying these conditions is an admissible extension *of Ω.*

The horizon condition serves to exclude examples which hide Ω inside an arbitrarily small neck, which would force the infimum to zero. This is a refinement [30] of the original definition [4], which prohibited horizons altogether.

Clearly $m_{QL}(\Omega)$ is well-defined and finite, once the region Ω admits just one admissible extension. The PMT with horizon boundary implies non-negativity, and monotonicity follows directly. Strict positivity of m_{QL} was established in [30], with the slightly weaker rigidity conclusion that if $m_{QL}(\Omega) = 0$ then Ω is locally flat. Agreement with the spherical mass, and the ADM limit condition, follows also from [30]. Bray's results imply that $m_{QL}(\Omega)$ agrees with the black hole mass in the limit as Ω shrinks down to a black hole. In addition, $m_{QL}(\Omega) \leq m_{ADM}(M)$ for any admissible extension M, so m_{QL} is the optimal quasi-local mass definition with respect to this condition.

The optimal form of the horizon condition remains conjectural. Bray has suggested an alternative condition, that Ω be a "strictly minimizing hull" [30] in M, so $\Sigma = \partial\Omega$ has the least area amongst all enclosing surfaces in the exterior. In this case we say Σ is *outer minimizing*, and denote by $\tilde{m}_{QL}(\Omega)$ the quasilocal mass function defined by restricting admissible extensions to those M in which Σ is outer minimizing. For this modified definition the Penrose inequality [30, 12] applies to show that if $\partial\Omega$ embeds into the Schwarzschild 3-manifold with the same induced metric and mean curvature (cf. (4.1), (4.2)) and encloses the horizon, then $m_{QL}(\Omega) = \text{M}$. It is not clear how to establish this natural result for the unmodified definition $m_{QL}(\Omega)$.

4. Static metrics

Although in many respects the definition of m_{QL} is quite satisfactory, it is not constructive, and thus it is important to determine computational methods. The key is the following [4]

Conjecture 5 *The infimum in m_{QL} is realised by a 3-metric agreeing with Ω in the interior,* static *(1.4) in the exterior region, and such that the metric is Lipschitz-continuous across the matching surface Σ, and the mean curvatures of the two sides agree along Σ.*

A similar conjecture for the space-time generalisation of the quasi-local mass, asserts that the exterior metric is *stationary*, ie. admits a timelike Killing field [4, 7].

As motivation for this conjecture, note first that if $R(g) > 0$ in some region, then a conformal factor ϕ can be found such that $\phi^4 g$ has less mass and $R(\phi^4 g) \geq 0$. Thus a mass-minimizing metric for (4), if such a metric exists, must have vanishing scalar curvature. Now if the linearization $DR(g)h = \delta_g\delta_g h - \Delta tr_g h - Ric \cdot h$ is surjective then g admits a variation which produces positive scalar curvature. The formal obstruction to surjectivity is non-trivial $\ker DR(g)^*$, which leads to the static metric equations (1.4). Corvino [17] shows that if $\ker DR(g)^*$ is trivial in $U \subset M$ then there are compactly supported metric variations in U which increase the scalar curvature. This gives

Theorem 6 *If (M, g) realizes the infimum in Definition 4, then there is a $V \in C^\infty(M\backslash\Omega)$ such that g, V satisfy the static metric equations (1.4) in $M\backslash\Omega$.*

This suggests a computational algorithm for determining $m_{QL}(\Omega)$: find an asymptotically flat static metric with boundary geometry matching that of $\partial\Omega$. To determine the appropriate boundary conditions, recall the second variation formula for the area of the leaves of a foliation labelled by r:

$$R(g) = 2D_n H - |II|^2 - H^2 + 2K - 2\lambda^{-1}\Delta_r\lambda \tag{4.1}$$

where II, H, K are respectively the second fundamental form, mean curvature and Gauss curvature of the leaves, λ is the lapse function, $n = \lambda^{-1}\partial_r$ is the normal vector and Δ_r is the Laplacian on the leaves. Our conventions give $H = -D_n(\log\sqrt{\det g_r})$ where g_r is the volume element of the leaves. This shows that $R(g)$ will be defined distributionally across a matching surface as a bounded function if

$$\begin{aligned} g|_{T\partial\Omega} &= g|_{T\Sigma}, \\ H_{\partial\Omega} &= H_\Sigma. \end{aligned} \tag{4.2}$$

Conjecture 7 *(Ω, g) determines a unique static asymptotically flat manifold (S, g) with boundary $\Sigma \simeq \partial\Omega$ satisfying (4.2).*

If true, this would give a prime candidate for the minimal mass extension. It is known (Pengzi Miao, private communication) that the boundary conditions (4.2) are elliptic for (1.4).

It is tempting to conjecture that mass-minimizing sequences for m_{QL} should converge to a static metric. For example, [3, Theorem 5.2] shows that a sequence of metrics g_k, close in the weighted Sobolev space $W^{2,q}_{-\tau}$, $q > 3, \tau > 1/2$, to the flat metric δ on $\mathbb{R}^3$ and such that $m_{ADM}(g_k) \to 0$, converges strongly to δ in $W^{1,2}$. Similar results, under rather different size conditions, are given in [21], and a discussion of the general "weak compactness" conjecture may be found in [30].

5. Estimating quasi-local mass

To estimate m_{QL} from above, it suffices to construct admissible extensions — metrics with non-negative scalar curvature and satisfying (4.2). These boundary conditions exclude the usual conformal method. Instead, metrics in *quasi-spherical* form [6]

$$g = u^2\,dr^2 + (r\,d\vartheta + \beta^1 dr)^2 + (r\sin\vartheta\,d\varphi + \beta^2 dr)^2 \tag{5.1}$$

satisfy a parabolic equation for u on S^2 evolving in the radial direction, when $R(g) = 0$, with β^1, β^2 freely specifiable. Since the metric 2-spheres S^2_r have mean curvature $H_r = (2 - \mathrm{div}_{S^2}\beta)/ur > 0$, (5.1) provides admissible extensions for $\partial\Omega = S^2_r$ with mean curvature $H > 0$. The underlying parabolic equation derives from (4.1), and has been generalized to non-spherical foliations in [49]. As an application, choosing $\beta = 0$ we can show

Theorem 8 *Suppose $\partial\Omega = S^2_r$ metrically, with $H \geq 0$. Then*

$$m_{QL}(\Omega) \leq \tfrac{1}{2}r(1 - \tfrac{1}{4}r^2 \min_{\partial\Omega} H^2). \tag{5.2}$$

This bound is sharp when Ω is a flat ball or a Schwarzschild horizon.

Finding lower bounds for $m_{QL}(\Omega)$ is more difficult. Bray's definition of *inner mass* [12, p243] gives a lower bound, but for $\tilde{m}_{QL}(\Omega)$. The difficulty here as above lies in showing that a horizon inside Ω remains outermost when the inner region is glued to a general exterior region $M_{\text{ext}} \subset M \in \mathcal{PM}(\Omega)$. This follows easily when $\Sigma = \partial\Omega$ is outer-minimizing in M_{ext}, as guaranteed by the definition for $\tilde{m}_{QL}(\Omega)$.

On physical grounds one expects that if "too much" matter is compressed into region which is "too small", then a black hole must be present. The geometric challenge lies in making this heuristic statement precise, and the only result in this direction has been [47], which gives quantitative measures which guarantee the existence of a black hole. An observation by Walter Simon (private communication) is thus very interesting: if $m_{QL}(\Omega) = 1$ (say) and Ω embeds isometrically into a complete asymptotically flat manifold M without boundary and with non-negative scalar curvature, and such that $m_{ADM}(M) < 1$, then M must have a horizon. This reinforces the importance of finding good lower bounds for m_{QL}, since the existence of a horizon in a similar situation with $\tilde{m}_{QL}$ does not follow.

References

[1] R. Arnowitt, S. Deser, and C. Misner. Coordinate invariance and energy expressions in general relativity. *Phys. Rev.*, 122:997–1006, 1961.

[2] C. Bär. Lower eigenvalue estimates for Dirac operators. *Math. Ann.*, 293:39–46, 1992.

[3] R. Bartnik. The mass of an asymptotically flat manifold. *Comm. Pure Appl. Math.*, 39:661–693, 1986.

[4] R. Bartnik. New definition of quasilocal mass. *Phys. Rev. Lett.*, 62(20):2346–2348, May 1989.

[5] R. Bartnik. The regularity of variational maximal surfaces. *Acta Math.*, 1989.

[6] R. Bartnik. Quasi-spherical metrics and prescribed scalar curvature. *J. Diff. Geom.*, 37:31–71, 1993.

[7] R. Bartnik. Energy in general relativity. In Shing-Tung Yau, editor, *Tsing Hua Lectures on Analysis and Geometry*, pages 5–28. International Press, 1997.

[8] R. Bartnik and P. Chruściel. On spectral boundary conditions for Dirac-type equations. preprint, 2002.

[9] G. Bergqvist. Quasilocal mass for event horizons. *Class. Quant. Grav.*, 9:1753–1768, 1992.

[10] G. Bergqvist. Positivity and definitions of mass. *Class. Quant. Gravity*, 9:1917–1922, 1992.

[11] D. H. Bernstein and K. P. Tod. Penrose's quasilocal mass in a numerically computed space-time. *Phys. Rev.*, D49:2808–2819, 1994.

[12] H. Bray. Proof of the Riemannian Penrose inequality using the positive mass theorem. *J. Diff. Geom.*, 59:177–267, 2001.

[13] G. Bunting and A. K. M. Masood ul alam. Non-existence of multiple black holes in asymptotically Euclidean static vacuum space-time. *Gen. Rel. Grav.*, 19:147, 1987.

[14] D. Christodoulou and S.-T. Yau. Some remarks on the quasi-local mass. In J. Isenberg, editor, *Mathematics and General Relativity*, Contemporary Mathematics. American Math. Society, 1988.

[15] P. T. Chruściel. Boundary conditions at spatial infinity from a Hamiltonian point of view. In P. G. Bergmann and V. de Sabbata, editors, *Topological properties and global structure of space-time.* Plenum, New York, 1986.

[16] P. T. Chruściel and G. Nagy. The mass of spacelike hypersurfaces in asymptotically andi-de Sitter spacetimes. *Adv. Theor. Math. Phys.*, 5, 2001.

[17] J. Corvino. Scalar curvature deformation and a gluing construction for the Einstein constraint equations. *Commun. Math. Phys.*, 214:137–189, 2000.

[18] A. Dimakis and F. Müller-Hoissen. Spinor fields and the positive enegy theorem. *Class. Quantum Grav.*, 7:283–295, 1990.

[19] A. J. Dougan and L. J. Mason. Quasi-local mass constructions with positive gravitational energy. *Phys. Rev. Lett.*, 67:2119–2123, 1991.

[20] D. M. Eardley. Global problems in numerical relativity. In L. Smarr, editor, *Sources of Gravitational Radiation*, pages 127–138. Cambridge UP, 1979.

[21] F. Finster and I. Kath. Curvature estimates in asymptotically flat manifolds of positive scalar curvature. MPI Leipzig preprint, 2001.

[22] G. T. Horowitz G. W. Gibbons, S. W. Hawking and M. J. Perry. Positive mass theorems for black holes. *Commun. Math. Phys.*, 88:295–308, 1983.

[23] R. Geroch. Energy extraction. *Ann. N.Y. Acd. Sci.*, 224:108–117, 1973.

[24] G. W. Gibbons. The isoperimetric and Bogomolny inequalities for black holes. In T. Willmore and N. Hitchin, editors, *Global Riemannian Geometry*, chapter 6, pages 194–202. Ellis Harwood, Chichester, April 1984.

[25] S. W. Hawking. Gravitational radiation in an expanding universe. *J. Math. Phys.*, 9:598–604, 1968.

[26] S. W. Hawking and G. R. Ellis. *The large-scale structure of spacetime.* Cambridge UP, Cambridge, 1973.

[27] M. Herzlich. A Penrose-like inequality for the mass on Riemannian asymptotically flat manifolds. *Commun. Math. Phys.*, 188:121–133, 1997.

[28] O. Hijazi. A conformal lower bound for the smallest eigenvalue of the Dirac operator and Killing spinors. *Commun. Math. Phys.*, 104:151–162, 1986.

[29] G. Huisken and T. Ilmanen. The Riemannian Penrose inequality. *Int. Math. Res. Not.*, 20:1045–1058, 1997.

[30] G. Huisken and T. Ilmanen. The inverse mean curvature flow and the Riemannian Penrose inequality. *J. Diff. Geom.*, 59:353–438, 2001.

[31] W. Israel. Event horizons in static electrovac space-times. *Commun. Math. Phys.*, 8:245–260, 1968.

[32] P. S. Jang. On positivity of mass for black hole space-times. *Commun. Math. Phys.*, 69:257–266, 1979.

[33] J. Kijowski. Unconstrained degrees of freedom of gravitational field and the positivity of gravitational energy. In *Gravitation, Geometry and Relativistic*

Physics, LNP 212. Springer, 1984.
[34] J. Lee and T. Parker. The Yamabe problem. *Bull. AMS*, 17:37–81, 1987.
[35] A. Lichnerowicz. Spineurs harmonique. *C.R. Acad. Sci. Paris Sér. A-B*, 257:7–9, 1963.
[36] J. Lohkamp. Scalar curvature and hammocks. *Math. Ann.*, 313:385–407, 1999.
[37] N. Ó Murchadha. Total energy momentum in general relativity. *J. Math. Phys.*, 27:2111–2128, 1986.
[38] J. M. Nester. The gravitational Hamiltonian. In F. J. Flaherty, editor, *Asymptotic behaviour of mass and space-time geometry (Oregon 1983)*, Lecture Notes in Physics 212, pages 155–163. Springer Verlag, 1984.
[39] J. M. Nester. A gauge condition for orthonormal three-frames. *J. Math. Phys.*, 30:624–626, 1988.
[40] R. Penrose. Naked singularities. *Ann. N. Y. Acad. Sci.*, 224:125–134, 1973.
[41] R. Penrose. Quasi-local mass and angular momentum in general relativity. *Proc. Roy. Soc. Lond. A*, 381:53–63, 1982.
[42] E. Woolgar R. Penrose, R.D. Sorkin. A positive mass theorem based on the focusing and retardation of null geodesics. grqc/9301015, 1993.
[43] D. C. Robinson. A simple proof of the generalisation of Israel's theorem. *Gen. Rel. Grav.*, 8:695–698, 1977.
[44] R. Schoen. Conformal deformation of a Riemannian metric to constant scalar curvature. *J. Diff. Geom.*, 20:479–495, 1984.
[45] R. Schoen and S.-T. Yau. Proof of the positive mass theorem. *Comm. Math. Phys.*, 65:45–76, 1979.
[46] R. Schoen and S.-T. Yau. Proof of the positive mass theorem II. *Comm. Math. Phys.*, 79:231–260, 1981.
[47] R. Schoen and S.-T. Yau. The existence of a black hole due to the condensation of matter. *Comm. Math. Phys.*, 90:575–579, 1983.
[48] E. Schrödinger. Diracsches Elektron im Schwerfeld. *Preuss. Akad. Wiss. Phys.-Math.*, 11:436–460, 1932.
[49] B. Smith and G. Weinstein. On the connectedness of the space of initial data for the Einstein equations. *Electron. Res. Announc. Amer. Math. Soc.*, 6:52–63, 2000.
[50] K. P. Tod. Some examples of Penrose's quasi-local mass construction. *Proc. Roy. Soc. Lond. A*, 388:457–477, 1983.
[51] X. Wang. The mass of asymptotically hyperbolic manifolds. *J. Diff. Geom.*, 57:273–300, 2001.
[52] E. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Comm. Pure and Appl. Math.*, 13, 1960.
[53] E. Witten. A simple proof of the positive energy theorem. *Comm. Math. Phys.*, 80:381–402, 1981.

ICM 2002 · Vol. II · 241–255

Geometry of Symplectic Intersections

P. Biran*

Abstract

In this paper we survey several intersection and non-intersection phenomena appearing in the realm of symplectic topology. We discuss their implications and finally outline some new relations of the subject to algebraic geometry.

2000 Mathematics Subject Classification: 53D35, 53D40, 14D06, 14E25.
Keywords and Phrases: Symplectic, Lagrangian, Algebraic variety.

1. Introduction

Symplectic geometry exhibits a range of intersection phenomena that cannot be predicted nor explained on the level of pure topology or differential geometry. The main players in this game are certain pairs of subspaces (e.g. Lagrangian submanifolds, domains, or a mixture of both) whose mutual intersections cannot be removed (or reduced) via the group of Hamiltonian or symplectic diffeomorphisms. The very first examples of such phenomena were conjectures by Arnold in the 1960's, and eventually established and further explored by Gromov, Floer and others starting from the mid 1980s.

The first part of the paper will survey several intersection phenomena and the mathematical tools leading to their discovery. We shall not attempt to present the most general results and since the literature is vast the exposition will be far from complete. Rather we shall concentrate on various intersection phenomena trying to understand their nature and whether there is any relations between them.

The second part is dedicated to "non-intersections", namely to situations where the principles of symplectic intersections break down. In the case of Lagrangian submanifolds this absence of intersections is reflected in the vanishing of a symplectic invariant called Floer homology. This vanishing when interpreted algebraically leads to restrictions on the topology of Lagrangian submanifolds. As a byproduct we shall explain how these restrictions can be used to study some problems in algebraic geometry concerning hyperplane sections and degenerations.

*School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel. Email: biran@math.tau.ac.il

2. Various intersection phenomena

In this section we shall make a brief tour through the zoo of symplectic intersections, encountering three different species.

Before we start let us recall two important notions from symplectic geometry. Let (M,ω) be a symplectic manifold. A submanifold $L \subset M$ is called *Lagrangian* if $\dim L = \frac{1}{2}\dim M$ and ω vanishes on $T(L)$. From now on we assume all Lagrangian submanifolds to be closed. The second notion is of *Hamiltonian isotopies.* An isotopy of diffeomorphisms $\{h_t : M \to M\}_{0\leq t\leq 1}$, starting with $h_0 = \mathrm{id}$ is called Hamiltonian if the (time-dependent) vector field ξ_t generating it satisfies that the 1-forms $i_{\xi_t}\omega$ are *exact* for all $0 \leq t \leq 1$. Note that Hamiltonian isotopies preserve the symplectic structure: $h_t^*\omega = \omega$ for all t. Finally, two subsets $A, B \subset M$ are said to be Hamiltonianly isotopic if there exists a Hamiltonian isotopy h_t such that $h_1(A) = B$. We refer the reader to [28] for the foundations of symplectic geometry.

2.1. Lagrangians intersect Lagrangians

The most fundamental *Lagrangian intersection* phenomenon occurs in cotangent bundles. Let X be a closed manifold and $T^*(X)$ be its cotangent bundle endowed with the canonical symplectic structure $\omega_{\mathrm{can}} = \sum dp_i \wedge dq_i$. Denote by $\lambda_{\mathrm{can}} = \sum p_i dq_i$ the Liouville form (so that $\omega_{\mathrm{can}} = d\lambda_{\mathrm{can}}$). Recall that a Lagrangian submanifold $L \subset T^*(X)$ is called exact if the restriction $\lambda_{\mathrm{can}}|_{T(L)}$ is exact. Note that the property of exactness is preserved by Hamiltonian isotopies. Denote by $O_X \subset T^*(X)$ the zero-section. The following theorem was proved by Gromov in [22]:

Theorem A. *Let $L \subset T^*(X)$ be an exact Lagrangian submanifold. Then:*
1) For every Lagrangian L' which is Hamiltonianly isotopic to L we have $L \cap L' \neq \emptyset$.
2) $L \cap O_X \neq \emptyset$. In particular, L cannot be separated from the zero-section by any Hamiltonian isotopy.

If one assumes L to be a Hamiltonian image of the zero-section a more quantitative version of Theorem A holds:

Theorem B. *Let $L \subset T^*(X)$ be a Lagrangian submanifold which is Hamiltonianly isotopic to the zero-section and intersects it transversely. Then*

$$\# L \cap O_X \geq \sum_{j=0}^{\dim X} b_j(X),$$

where $b_j(X)$ are the Betti numbers of X.

Chronologically Theorem B preceded Theorem A. It was conjectured by Arnold (see [3] for the history), first proved for $X = \mathbb{T}^n$ by Chaperon [12] and generalized to all cotangent bundles by Hofer [23] and by Laudenbach and Sikorav [26]. Now a days it can be viewed as a special case of Floer theory (see Section 2.4 below).

Note that the intersections described by both theorems above cannot in general be understood on a purely topological level. Indeed, in general topology predicts less than $\sum b_j(X)$ intersection points, and sometimes even none. Finally, note that in general the statement of Theorem B fails if one assumes L to be only symplectically isotopic to O_X, as the example $X = \mathbb{T}^n$ shows.

2.2. Balls intersect balls

Denote by $B^{2n}(R)$ the closed Euclidean ball of radius R, endowed with the standard symplectic structure induced from $\mathbb{R}^{2n}$. Denote by $\mathbb{C}P^n$ the complex projective space, endowed with its standard Kähler form σ, normalized so that $\int_{\mathbb{C}P^1} \sigma = \pi$. The following obstruction for *symplectic packing* was discovered by Gromov [22]:

Theorem C. *Let M be either $B^{2n}(1)$ or $\mathbb{C}P^n$. Let $B_{\varphi_1}, B_{\varphi_2} \subset M$ be the images of two symplectic embeddings $\varphi_1 : B^{2n}(R_1) \to M$, $\varphi_2 : B^{2n}(R_2) \to M$. If $R_1^2 + R_2^2 \geq 1$ then $B_{\varphi_1} \cap B_{\varphi_2} \neq \emptyset$.*

Since symplectic embeddings are also volume preserving there is an obvious volume obstruction for having $B_{\varphi_1} \cap B_{\varphi_2} = \emptyset$. However, volume considerations predict an intersection only if $R_1^{2n} + R_2^{2n} \geq 1$ (moreover for volume preserving embeddings the latter inequality is sharp).

When one considers embeddings of several balls things become more complicated and interesting. Here results are currently available only in dimension 4.

Theorem D. *Let M be either $B^4(1)$ or $\mathbb{C}P^2$, and let $B_{\varphi_1}, \ldots, B_{\varphi_N} \subset M$ be the images of symplectic embeddings $\varphi_k : B^4(R) \to M$, $k = 1, \ldots, N$, of N balls of the same radius R. Then there exist $i \neq j$ such that $B_{\varphi_i} \cap B_{\varphi_j} \neq \emptyset$ in each of the following cases:*

1. *$N = 2$ or 3 and $R^2 \geq 1/2$.*
2. *$N = 5$ or 6 and $R^2 \geq 2/5$.*
3. *$N = 7$ and $R^2 \geq 3/8$.*
4. *$N = 8$ and $R^2 \geq 6/17$.*

Moreover all the above inequalities are sharp in the sense that in each case if the inequality on R is not satisfied then there exist symplectic embeddings $\varphi_1, \ldots, \varphi_N$ as above with disjoint images $B_{\varphi_1}, \ldots, B_{\varphi_N} \subset M$.

Statement 2 for $N = 5$ was proved by Gromov [22]. The rest was established by McDuff and Polterovich [27]. Let us mention that for $N = 4$ and any $N \geq 9$ this intersection phenomenon completely disappears in the sense that an arbitrarily large portion of the volume of M can be filled by a disjoint union of N equal balls (see [27] for $N = 4$ and $N = k^2$, and [5, 6] for the remaining cases).

2.3. Balls intersect Lagrangians

It turns out that there exist (symplectically) irremovable intersections also between contractible domains (e.g. balls) and Lagrangian submanifolds.

Denote by $\mathbb{R}P^n \subset \mathbb{C}P^n$ the Lagrangian n-dimensional real projective space (embedded as the fixed point set of the standard conjugation of $\mathbb{C}P^n$). The following was proved in [7]:

Theorem E. *Let $B_\varphi \subset \mathbb{C}P^n$ be the image of a symplectic embedding $\varphi : B^{2n}(R) \to \mathbb{C}P^n$. If $R^2 \geq 1/2$ then $B_\varphi \cap \mathbb{R}P^n \neq \emptyset$. Moreover the inequality is sharp, namely for every $R^2 < 1/2$ there exists a symplectic embedding $\varphi : B^{2n}(R) \to \mathbb{C}P^n$ whose image avoids $\mathbb{R}P^n$.*

In fact this pattern of intersections occurs in a wide class of examples (see [7]):

Theorem E'. *Let (M,ω) be a closed Kähler manifold with $[\omega] \in H^2(M;\mathbb{Q})$ and $\pi_2(M) = 0$. Then for every $\epsilon > 0$ there exists a Lagrangian CW-complex $\Delta_\epsilon \subset (M,\omega)$ with the following property: every symplectic embedding $\varphi : B^{2n}(\epsilon) \to (M,\omega)$ must satisfy* $\text{Image}(\varphi) \cap \Delta_\epsilon \neq \emptyset$.

By a Lagrangian CW-complex we mean a subspace $\Delta_\epsilon \subset M$ which topologically is a CW-complex and the interior of each of its cells is a smoothly embedded disc of M on which ω vanishes.

2.4. Methods for studying intersections

Lagrangian intersections. The first systematic study of Lagrangian intersections was based on the theory of generating function [12, 26] (an equivalent theory was independently developed in contact geometry [13]). Gromov's theory of pseudo-holomorphic curves [22] gave rise to an alternative approach which culminated in what is now called Floer theory. Each of these theories has its own advantage. Floer theory works in larger generality and seems to have a richer algebraic structure, on the other hand the theory of generating functions leads in some cases to sharper results (see [20]).

Since Floer theory will appear in the sequel, let us outline a few facts about it (the reader is referred to the works of Floer [16] and of Oh [29, 30] for details). Let (M,ω) be a symplectic manifold and $L_0, L_1 \subset (M,\omega)$ two Lagrangian submanifolds. In "ideal" situations Floer theory assigns to this data an invariant $HF(L_0,L_1)$. This is a $\mathbb{Z}_2$-vector space obtained through an infinite dimensional version of Morse-Novikov homology performed on the space of paths connecting L_0 to L_1. The result of this theory is a chain complex $CF(L_0,L_1)$ whose underlying vector space is generated by the intersection points $L_0 \cap L_1$ (one perturbs L_0, L_1 so their intersection becomes transverse). The homology of this complex $HF(L_0,L_1)$ is called the Floer homology of the pair (L_0,L_1). The most important feature of $HF(L_0,L_1)$ is its invariance under Hamiltonian isotopies: if L_0', L_1' are Hamiltonianly isotopic to L_0, L_1 respectively, then $HF(L_0',L_1') \cong HF(L_0,L_1)$. From this point of view $HF(L_0,L_1)$ can be regarded as a quantitative obstruction for Hamiltonianly separating L_0 from L_1. Indeed, the rank of $HF(L_0,L_1)$ is a lower bound on the number of intersection points of any pair of transversally intersecting Lagrangians L_0', L_1' in the Hamiltonian deformation classes of L_0, L_1 respectively.

Let us explain the "ideal situations" in which Floer homology is defined. First of all there are restrictions on M: due to analytic difficulties manifolds are required to be either closed or to have symplectically convex ends (e.g. $\mathbb{C}^n$, cotangent bundles or any Stein manifold). More serious restrictions are posed on the Lagrangians. For simplicity we describe them only for the case when L_1 is Hamiltonianly isotopic to L_0. From now on we shall write $L = L_0$ and $L' = L_1$. In Floer's original setting [16] the theory was defined under the assumption that the homomorphism $A_\omega : \pi_2(M,L) \to \mathbb{R}$, defined by $D \mapsto \int_D \omega$, vanishes. The reason for this comes from the construction of the differential of the Floer complex: the main obstruction for defining a meaningful differential turns out to be existence of holomorphic discs with boundary on L or L'. These discs appear as a source of non-compactness of the space of solutions of the PDEs involved in the construction. Since holomorphic

discs must have positive symplectic area the assumption $A_\omega = 0$ rules out their existence. Under this assumption Floer defined $HF(L, L')$ and proved its invariance under Hamiltonian isotopies. Moreover he showed that $HF(L, L)$ is isomorphic to the singular cohomology $H^*(L; \mathbb{Z}_2)$ of L. This together with the invariance give:

Theorem F. *Let (M, ω) be a symplectic manifold, either compact or with symplectically convex ends. Let $L \subset (M, \omega)$ be a Lagrangian submanifold with $A_\omega = 0$. Then for every Lagrangian L' which is Hamiltonianly isotopic to L and intersects L transversally we have: $\# L \cap L' \geq \mathrm{rank} HF(L, L') = \mathrm{rank} H^*(L; \mathbb{Z}_2)$. In particular L cannot be separated from itself by a Hamiltonian isotopy.*

Floer theory was extended by Oh [30] to cases when $A_\omega \neq 0$. There are two assumptions needed for this extension to work: the Maslov homomorphism $\mu : \pi_2(M, L) \to \mathbb{Z}$ should be positively proportional to A_ω (such Lagrangians are called monotone). The second assumption is that the positive generator N_L of the subgroup $\mathrm{Image}\, \mu \subset \mathbb{Z}$ is at least 2. In this setting Oh defined $HF(L, L')$ and proved its invariance under Hamiltonian isotopies. It is however no longer true in general that $HF(L, L)$ is isomorphic to $H^*(L; \mathbb{Z}_2)$. Still, Oh proved [29] that $HF(L, L)$ is related to $H^*(L; \mathbb{Z}_2)$ through a spectral sequence. Recently the theory was considerably generalized by Fukaya, Oh, Ohta and Ono [21].

Intersections of balls. Theorems C and D were obtained using Gromov's theory of pseudo-holomorphic curves. The hard-core of the proofs consists of existence of pseudo-holomorphic curves of specified degrees that pass through a prescribed number of points in the manifold (see [22, 27]) for the details). From a more modern perspective it can be viewed as an early application of Gromov-Witten invariants.

Finally, Theorems E and E' are proved by a decomposition technique introduced in [7] which enables to decompose symplectic manifolds as a disjoint union of a symplectic disc bundle and a Lagrangian CW-complex. A variation on the proof of Gromov's non-squeezing theorem [22] gives an upper bound on the radius of a symplectic ball that can be squeezed inside that disc bundle. Hence, a larger ball must always intersect this CW-complex. For $M = \mathbb{C}P^n$, the corresponding CW-complex turns out to be a smooth copy of $\mathbb{R}P^n$. See [7] for the details.

3. Some questions and speculations

Cotangent bundles. The following questions show that even in the case of cotangent bundles the most fundamental invariants are not completely understood.

1. Let $L \subset T^*(X)$ be an exact Lagrangian (not necessarily Hamiltonianly isotopic to O_X). By Theorem A, $L \cap O_X \neq \emptyset$. Is it true that $HF(L, O_X) \neq 0$?
2. Let $L_0, L_1 \subset T^*(X)$ be two exact Lagrangians (again, not necessarily Hamiltonianly isotopic neither to O_X nor to each other). Is it true that $L_0 \cap L_1 \neq \emptyset$? Is it true that $HF(L_0, L_1) \neq 0$?

These questions are of a theoretical importance, since the zero section and its Hamiltonian images are the only known examples of exact Lagrangians in $T^*(X)$.

Symplectic packing. Lack of tools (or new ideas) prevent us from understanding

symplectic packings in dimension higher than 4. The only packing obstructions known in these dimensions are described in Theorem C. Note that $\mathbb{C}P^n$ admits full packing by $N = k^n$ equal balls [27], but it is unclear what happens for other values of N. In view of this and Theorem C, the first unknown case (for $n \geq 3$) is of $N = 2^n + 1$ equal balls.

The situation in dimension 4 is only slightly better. Except of $\mathbb{C}P^2$ and a few other rational surfaces no packing obstructions are known. It is known that for every symplectic 4-manifold (M, ω) with $[\omega] \in H^2(M; \mathbb{Q})$ packing obstruction (for equal balls) disappear once the number of balls is large enough (see [6]), but nothing is known when the number of balls is small. In fact even the case of one ball is poorly understood (namely, what is the maximal radius of a ball that can be symplectically embedded in M). The reason here is that the methods yielding packing obstructions strongly rely on the geometry of algebraic and pseudo-holomorphic curves in the manifold. The problem is that most symplectic manifolds have very few (or none at all) J-holomorphic curves for a generic choice of the almost complex structure. Thus, even in dimension 4 it is unknown whether or not packing obstructions is a phenomenon particular to a sporadic class of manifolds such as $\mathbb{C}P^2$.

Is everything Lagrangian? Weinstein's famous saying could be relevant for the intersection described in Theorems C,D and E. In other words, it could be that these intersections are in fact Lagrangian intersections under disguise. To be more concrete, let $\frac{1}{2} < R^2 < \frac{1}{2^{1/n}}$ and consider a Lagrangian L_R lying on the boundary $\partial B^{2n}(R)$. Is it possible to Hamiltonianly separate L_R from itself inside $B^{2n}(1)$?

If we can find a Lagrangian L_R for which the answer is negative then this would strongly indicate that Theorem C is in fact a Lagrangian intersections result. Namely it would imply Theorem C for $R_1 = R_2$ under the additional assumption that φ_1, φ_2 are symplectically isotopic. A good candidate for L_R seems to be the split torus $\partial B^2(\sqrt{R/n}) \times \cdots \times \partial B^2(\sqrt{R/n}) \subset \partial B^{2n}(R)$, but one could try other Lagrangians as well.

Attempts to approach this question with traditional Floer homology fail. The reason is that Floer homology is blind to sizes: both to the "size" of the Lagrangian L_R as well as to the "size" of the domain in which we work $B^{2n}(1)$. Indeed it is easy to see that $HF(L_R, L_R)$ whether computed inside $B^{2n}(1)$ or in $\mathbb{R}^{2n}$ is the same, hence vanishes. The meaning of "sizes" can be made precise: the size of L_R is encoded in its Liouville class, and the size of $B^{2n}(1)$ could be encoded here by the action spectrum of its boundary.

It would be interesting to try a mixture of symplectic field theory [19] with Floer homology. This would require a sophisticated counting of holomorphic discs with k punctures (for all $k \geq 0$), where the boundary of the discs goe to L_R and the punctures to periodic orbits on $\partial B^{2n}(1)$.

It is interesting to note that when the radii of the balls are not equal things become more complicated. Indeed suppose that $R_1^2 + R_2^2 > 1$ and consider two Lagrangian submanifolds L_{R_1}, L_{R_2} lying on the boundaries of the balls $B_{\varphi_1}, B_{\varphi_2}$. Then clearly L_{R_1} and L_{R_2} can be disjoint even though the balls $B_{\varphi_1}, B_{\varphi_2}$ do intersect (e.g. two concentric balls $B_{\varphi_1} \subset B_{\varphi_2}$, where $R_1 < R_2$). It would be interesting

to see to which extent this mutual position can be detected on the level of the Lagrangians L_{R_1} and L_{R_2} alone. Or, in more pictorial (but less mathematical) terms, do the Lagrangians L_{R_1} and L_{R_2} know that they lie one "inside" the other ?

Returning to the case of equal balls, if the above plan is feasible, it would be interesting to try similar approaches for more than two balls as described in Theorem D. A similar approach could be tried in the situation of Theorem E. Here one could expect an irremovable intersection between a Lagrangian submanifold $L_R \subset \partial B_\varphi$ and $\mathbb{R}P^n$.

Quantitative intersections. In contrast to the quantitative version of Lagrangian intersections given by Theorems B and F, Theorems C–E provide only existence of intersections. Is it possible to measure the size of these intersections ?

More concretely, consider two balls $B_{\varphi_1}, B_{\varphi_2} \subset B^{2n}(1)$ with $R_1^2 + R_2^2 > 1$ but with $R_1^{2n} + R_2^{2n} < 1$ (so that $\mathrm{Vol}(B_{\varphi_1}) + \mathrm{Vol}(B_{\varphi_2}) < 1$). Is it possible to bound from below the size of $B_{\varphi_1} \cap B_{\varphi_2}$?

It is not hard to see that volume is a wrong candidate for the size since for every $\epsilon > 0$ there exist two such balls with $\mathrm{Vol}(B_{\varphi_1} \cap B_{\varphi_2}) < \epsilon$. Symplectic capacities seem also to be inappropriate for this task. It could be that "size" should be replaced here by a kind of "complexity" or a trade-off between capacity and complexity: namely if the intersection has large capacity (e.g. when $B_{\varphi_1} \subset B_{\varphi_2}$) the complexity is low, and vice-versa. Note that in dimension 2 a possible notion of complexity of a set is the number of connected components of its interior.

A related problem is the following. Consider two symplectic balls $B_{\varphi_1}, B_{\varphi_2} \subset \mathbb{C}P^n$ of radii R_1, R_2, where $R_1^2 + R_2^2 = 1$. Assume further that $\mathrm{Int}\,(B_{\varphi_1}) \cap \mathrm{Int}\,(B_{\varphi_2}) = \emptyset$. Theorem C implies that the balls must intersect hence the intersection occurs on the boundaries: $\partial B_{\varphi_1} \cap \partial B_{\varphi_2} \neq \emptyset$. What can be said about the intersection $\partial B_{\varphi_1} \cap \partial B_{\varphi_2} \neq \emptyset$, in terms of size, dynamical properties etc. ?

It is easy to see that this intersection cannot be discrete. Moreover, an argument based on the work of Sullivan [37] shows that the intersection must contain at least one entire (closed) orbit of the characteristic foliation of the boundaries of the balls (see [33] for a discussion on this point). Looking at examples however suggests that the number of orbits in the intersection should be much larger.

The same problem can be considered also for (some of) the extremal cases described in Theorem D. Similarly one can study the intersection $\partial B_\varphi \cap \mathbb{R}P^n$ where $B_\varphi \subset \mathbb{C}P^n$ is a symplectic ball of radius $R^2 = 1/2$ whose interior is disjoint from $\mathbb{R}P^n$. It is likely that methods of symplectic field theory [19] could shed some light on this circle of problems.

Stable intersections. The problems described here come from Polterovich [32]. Let (M, ω) be a symplectic manifold and $A \subset M$ a subset. We say that A has the *Hamiltonian intersection property* if for every Hamiltonian diffeomorphism f we have $f(A) \cap A \neq \emptyset$. We say that A has the *stable Hamiltonian intersection property* if $O_{S^1} \times A \subset T^*(S^1) \times M$ has the Hamiltonian intersection property. Polterovich discovered in [32] that if there exists a subset $A \subset M$ with open non-empty complement and with the stable Hamiltonian property then the universal cover $\widetilde{\mathrm{Ham}}(M, \omega)$ of the

group of Hamiltonian diffeomorphisms has infinite diameter with respect to Hofer's metric. Note that when $\pi_1(\mathrm{Ham}(M,\omega))$ is finite the same holds also for $\mathrm{Ham}(M,\omega)$ itself. (See [32] for the details and references for other results on the diameter of Ham). This is applicable when (M,ω) contains a Lagrangian submanifold A with $HF(A,A) \neq 0$, since then $HF(O_{S^1} \times A, O_{S^1} \times A) = (\mathbb{Z}_2 \oplus \mathbb{Z}_2) \otimes HF(A,A) \neq 0$. For example, taking $A = \mathbb{R}P^n \subset \mathbb{C}P^n$ Polterovich proved that $\mathrm{diam}\,\widetilde{\mathrm{Ham}}(\mathbb{C}P^n) = \infty$ (for $n = 1,2$ the same holds for $diam\,\mathrm{Ham}(\mathbb{C}P^n)$).

In view of the above the following question seems natural: does every closed symplectic manifold contain a subset A with open non-empty complement and with the stable Hamiltonian intersection property ? Note that besides Lagrangian submanifolds (with $HF \neq 0$) no other stable Hamiltonian intersection phenomena are known. It would also be interesting to find out whether the intersections described in Theorems C,D,E and especially E' continue to hold after stabilization.

4. Intersections versus non-intersections

In contrast to cotangent bundles there are manifolds in which *every compact subset* can be separated from itself by a Hamiltonian isotopy. The simplest example is $\mathbb{C}^n$: indeed linear translations are Hamiltonian, and any compact subset can be translated away from itself. Clearly the same also holds for every symplectic manifold of the type $M \times \mathbb{C}$ by applying translations on the $\mathbb{C}$ factor. Note that manifolds of the type $M \times \mathbb{C}$ sometime appear in "disguised" forms (e.g. as subcritical Stein manifolds, see Cieliebak [14]).

The "non-intersections" property has quite strong consequences on the topology of Lagrangian submanifolds already in $\mathbb{C}^n$. Denote by ω_{std} the standard symplectic structure of $\mathbb{C}^n$ and let λ be any primitive of ω_{std}. Note that the restriction $\lambda|_{T(L)}$ of λ to any Lagrangian submanifold $L \subset \mathbb{C}^n$ is closed. The following was proved by Gromov in [22]:

Theorem G. *Let $L \subset \mathbb{C}^n$ be a Lagrangian submanifold. Then the restriction of λ to L is not exact. In particular $H^1(L;\mathbb{R}) \neq 0$.*

Indeed if λ were exact on L then $A_\omega : \pi_2(\mathbb{C}^n, L) \to \mathbb{R}$ must vanish, hence by Theorem F it is impossible to separate L from itself by a Hamiltonian isotopy. On the other hand, as discussed above, in $\mathbb{C}^n$ this is always possible. We thus get a contradiction. (Gromov's original proof is somewhat different, however a careful inspection shows it uses the failure of Lagrangian intersections in an indirect way). Arguments exploiting non-intersections were further used in clever ways by Lalonde and Sikorav [25] to obtain information on the topology of exact Lagrangians in cotangent bundles (see also Viterbo [42] for further results).

An important property of symplectic manifolds W having the "non-intersections" property is the following vanishing principle: *for every Lagrangian submanifold $L \subset W$ with well defined Floer homology we have $HF(L,L) = 0$.* Applying this vanishing to $\mathbb{C}^n$ yields restrictions on the possible Maslov class of Lagrangian submanifolds of $\mathbb{C}^n$. (Conjectures about the Maslov class due to Audin appear already in [1]. First results in this directions are due to Polterovich [31] and to Viterbo [41]. The interpretation in Floer-homological terms is due to Oh [29]. Generalizations

to other manifolds appear in [2] and [11]. Finally, consult [21] for recent results answering old questions on the Maslov class).

4.1. Lagrangian embeddings in closed manifolds

The ideas described above can be applied to obtain information on the topology of Lagrangian submanifolds of some closed manifolds. Note that in comparison to closed manifolds the case of $\mathbb{C}^n$ can be regarded as local (Darboux Theorem). Of course, "local" should by no means be interpreted as easy. On the contrary, characterization of manifolds that admit Lagrangian embeddings into $\mathbb{C}^n$ is completely out of reach with the currently available tools.

Below we shall deal with the "global" case, namely with Lagrangians in closed manifolds. One (coarse) way to "mod out" local Lagrangians is to restrict to Lagrangians L with $H_1(L;\mathbb{Z})$ zero or torsion (so that by Theorem G they cannot lie in a Darboux chart). The pattern arising in the theorems below is that under such assumptions in some closed symplectic manifolds we have homological uniqueness of Lagrangian submanifolds. Let us view some examples.

We start with $\mathbb{C}P^n$. It is known that a Lagrangian submanifold $L \subset \mathbb{C}P^n$ cannot have $H_1(L;\mathbb{Z}) = 0$ (see Seidel [39], see also [10] for an alternative proof). However, $L \subset \mathbb{C}P^n$ may have torsion $H_1(L;\mathbb{Z})$ as the example $\mathbb{R}P^n \subset \mathbb{C}P^n$ shows.

Theorem H. *Let $L \subset \mathbb{C}P^n$ be a Lagrangian submanifold with $H_1(L;\mathbb{Z})$ a 2-torsion group (namely, $2H_1(L;\mathbb{Z}) = 0$). Then:*

1. *$H^*(L;\mathbb{Z}_2) \cong H^*(\mathbb{R}P^n;\mathbb{Z}_2)$ as graded vector spaces.*
2. *Let $a \in H^2(\mathbb{C}P^n;\mathbb{Z}_2)$ be the generator. Then $a|_L \in H^2(L;\mathbb{Z}_2)$ generates the subalgebra $H^{\text{even}}(L;\mathbb{Z}_2)$. Moreover if n is even the isomorphism in 1 is of graded algebras.*

Statement 1 of the theorem was first proved by Seidel [39]. An alternative proof based on "non-intersections" can be found in [8]. Let us outline the main ideas from [8]. Consider $\mathbb{C}P^n$ as a hypersurface of $\mathbb{C}P^{n+1}$. Let U be a small tubular neighbourhood of $\mathbb{C}P^n$ inside $\mathbb{C}P^{n+1}$. The boundary ∂U looks like a circle bundle over $\mathbb{C}P^n$ (in this case it is just the Hopf fibration). Denote by $\Gamma_L \to L$ the restriction of this circle bundle to $L \subset \mathbb{C}P^n$. A local computation shows that U can be chosen so that $\Gamma_L \subset \mathbb{C}P^{n+1} \setminus \mathbb{C}P^n$ becomes a Lagrangian submanifold. (This procedure works whenever we have a symplectic manifold Σ embedded as a hyperplane section in some other symplectic manifolds M). The next observation is that $\Gamma_L \subset \mathbb{C}P^{n+1} \setminus \mathbb{C}P^n$ is monotone and moreover its minimal Maslov number N_{Γ_L} is the same as the one of L. Due to our assumptions on $H_1(L;\mathbb{Z})$ this number turns out to satisfy $N_{\Gamma_L} \geq n+1$. The crucial point now is that $HF(\Gamma_L, \Gamma_L) = 0$. Indeed, the symplectic manifold $\mathbb{C}P^{n+1} \setminus \mathbb{C}P^n$ can be completed to be $\mathbb{C}^{n+1}$ where Floer homology vanishes.

Having this vanishing we turn to an alternative computation of $HF(\Gamma_L, \Gamma_L)$. This computation is based on the theory developed by Oh [29] for monotone Lagrangian submanifolds. According to [29] Floer homology can be computed via a spectral sequence whose first stage is the singular cohomology of the Lagrangian. The minimal Maslov number has an influence both on the grading as well as on the

number of steps it takes the sequence to converge to HF. In our case we have a spectral sequence starting with $H^*(\Gamma_L;\mathbb{Z}_2)$ and converging to $HF(\Gamma_L,\Gamma_L)=0$. A computation through this process together with the information that $N_{\Gamma_L}\geq n+1$ makes it possible to completely recover $H^*(\Gamma_L;\mathbb{Z}_2)$. It turns out that $H^i(\Gamma_L;\mathbb{Z}_2)=\mathbb{Z}_2$ for $i=0,1,n$ and $n+1$, while $H^i(\Gamma_L;\mathbb{Z}_2)=0$ for all $1<i<n$. Going back from $H^*(\Gamma_L;\mathbb{Z}_2)$ to $H^*(L;\mathbb{Z}_2)$ is now done by the Gysin exact sequence of the circle bundle $\Gamma_L\to L$ and noting that the second Stiefel-Whitney class of this bundle is nothing but the restriction $a|_L$ of the generator $a\in H^2(\mathbb{C}P^n;\mathbb{Z}_2)$.

Summarizing the proof, there are three main ingredients:

1. Transforming the Lagrangian L into a related Lagrangian Γ_L living in a different manifold such that Γ_L can be Hamiltonianly separated from itself. Consequently we obtain $HF(\Gamma_L,\Gamma_L)=0$.
2. Relating $HF(\Gamma_L,\Gamma_L)$ to $H^*(\Gamma_L)$ via the theory of Floer homology (e.g. a spectral sequence).
3. Passing back from $H^*(\Gamma_L)$ to $H^*(L)$.

Similar ideas work in various other cases (see [8]). For example, consider $\mathbb{C}P^n\times\mathbb{C}P^n$. This manifold has Lagrangians with $H_1(L;\mathbb{Z})=0$, e.g. $\mathbb{C}P^n$ which can be embedded as the "anti-diagonal" $\{(z,w)\in\mathbb{C}P^n\times\mathbb{C}P^n | w=\overline{z}\}$.

Theorem I. *Let $L\subset\mathbb{C}P^n\times\mathbb{C}P^n$ be a Lagrangian with $H_1(L;\mathbb{Z})=0$. Then $H^*(L;\mathbb{Z}_2)\cong H^*(\mathbb{C}P^n;\mathbb{Z}_2)$, the isomorphism being of graded algebras.*

Another application of this circle of ideas is for Lagrangian spheres. Recently Lagrangian spheres have attracted special attention due to their relations to interesting symplectic automorphisms [38, 39] and to symplectic Lefschetz pencils [15].

Theorem J. *1) Let M be a closed symplectic manifold with $\pi_2(M)=0$, and denote by $m=\dim_{\mathbb{C}}M$ its complex dimension. If $M\times\mathbb{C}P^n$ (where $m,n\geq 1$) has a Lagrangian sphere then $m\equiv n+1 \pmod{2n+2}$.*

2) Let $M=\mathbb{C}P^n\times\mathbb{C}P^m$, $m+n\geq 3$, be endowed with the split symplectic form $(n+1)\sigma\oplus(m+1)\sigma$. If M has a Lagrangian sphere then $\gcd(n+1,m+1)=1$.

Let us remark that when $m=n+1$ any product of the form $\mathbb{C}P^n\times M$ (with $\dim_{\mathbb{C}}M=n+1$) indeed has a Lagrangian sphere, after a possible rescaling of the symplectic form on the M factor (see [2], [10]). We are not aware of any other examples, namely when $m\equiv n+1 \pmod{2n+2}$ but $m\neq n+1$).

5. Relations to algebraic geometry

The purpose of this section is to show how ideas from Section 4 are related to algebraic geometry. We shall not present new results here but rather try to outline a new direction in which symplectic methods can be used in algebraic geometry.

5.1. Hyperplane sections

Let Σ be a smooth projective variety. The classical Lefschetz theorem provides restrictions on smooth varieties X that may contain Σ as their hyperplane section. It was discovered by Sommese [35] that there exist projective varieties Σ that cannot be hyperplane sections (or even ample divisors) in *any* smooth variety X. For

example, Sommese proved that Abelian varieties of (complex) dimension ≥ 2 have this property (see [35] for more examples).

Let us outline an alternative approach to this problem using symplectic geometry. Let $X \subset \mathbb{C}P^N$ be a smooth variety. Denote by $X^\vee \subset (\mathbb{C}P^N)^*$ the dual variety (namely, the variety of all hyperplanes $H \in (\mathbb{C}P^N)^*$ that are non-transverse to X).

Theorem K. *Suppose that $\Sigma = X \cap H_0 \subset X$ is a smooth hyperplane section of X obtained from a projective embedding $X \subset \mathbb{C}P^N$. Then either Σ has a Lagrangian sphere (for the symplectic structure induced from $\mathbb{C}P^N$), or* $\operatorname{codim}_{\mathbb{C}}(X^\vee) > 1$.

Here is an outline of the proof. Suppose that $\operatorname{codim}_{\mathbb{C}}(X^\vee) = 1$. Choose a generic line $\ell \subset (\mathbb{C}P^N)^*$ intersecting $X^\vee$ transversely (and only at smooth points of $X^\vee$). Consider the pencil $\{X \cap H\}_{H \in \ell}$ parametrized by ℓ. Passing to the blow-up $\widetilde{X}$ of X along the base locus of the pencil we obtain a holomorphic map $\pi : \widetilde{X} \to \ell \approx \mathbb{C}P^1$. The critical values of π are in 1-1 correspondence with the point of $\ell \cap X^\vee$. Moreover, the fact that ℓ intersects $X^\vee$ transversely implies that π is a so called Lefschetz fibration, namely each critical point of π has non-degenerate (complex) Hessian (in other words, locally π looks like a holomorphic Morse function). The condition $\operatorname{codim}_{\mathbb{C}}(X^\vee) = 1$ ensures that $\ell \cap X^\vee \neq \emptyset$ hence at least one of the fibres of π is singular. Let X_0 be such a fibre and $p \in X_0$ a critical point of π. The important point now is that the vanishing cycle (corresponding to p) that lies in the nearby smooth fibre X_ϵ can be represented by a (smooth) Lagrangian sphere. By Moser argument all the smooth divisors in the linear system $\{X \cap H\}_{H \in (\mathbb{C}P^N)^*}$ are symplectomorphic. In particular Σ has a Lagrangian sphere too.

The existence of Lagrangian vanishing cycles was known folklorically for long time. Its importance to symplectic geometry was realized by Arnold [4], Donaldson [15] and by Seidel [38].

Theorem K can be applied as follows: given a smooth variety Σ, use methods of symplectic geometry to prove that Σ contains no Lagrangian spheres, say for any symplectic structure compatible with the complex structure of Σ. Then by Theorem K the only chance for Σ to be a hyperplane section is inside a variety X with "small dual", namely $\operatorname{codim}_{\mathbb{C}}(X^\vee) > 1$. Let us remark that smooth varieties $X \subset \mathbb{C}P^N$ with $\operatorname{codim}_{\mathbb{C}}(X^\vee) > 1$ are quite rare, and have very restricted geometry (see e.g. Zak [43] and Ein [17, 18]). Using the theory of "small dual varieties" we can either rule out this case or get strong restrictions on the pair (X, Σ).

Let us illustrate this on the example mentioned at the beginning of the section. Let Σ be an Abelian variety of complex dimension $n \geq 2$. Note that Σ cannot have a Lagrangian sphere for any Kähler form. Indeed, if Σ had such a sphere then the same would hold also for the universal cover of Σ which is symplectomorphic to $\mathbb{C}^n$. But this is impossible in view of Theorem G. Thus if Σ is a hyperplane section of $X \subset \mathbb{C}P^N$ then $\operatorname{codim}_{\mathbb{C}}(X^\vee) > 1$. It is well known [24] that in this case X must have rational curves (in fact lots of them). In particular $\pi_2(X) \neq 0$. By Lefschetz's theorem we get $\pi_2(\Sigma) \neq 0$. But this is impossible since Σ is an Abelian variety. We therefore conclude that Σ cannot be a hyperplane section in *any* smooth variety X.

An analogous (though symplectically more involved) argument should apply also to any algebraic variety Σ with $c_1 = 0$ and $b_1(\Sigma) \neq 0$ (see [9]). An application of more refined symplectic tools (e.g. methods described in Section 4.1 above) can

be used to obtain many more examples.

Here is another typical application: let C be a projective curve of genus > 0. It was observed by Silva [34] that $C \times \mathbb{C}P^n$ can be realized as a hyperplane section in various smooth varieties. Note that by Theorem J, $C \times \mathbb{C}P^n$ cannot have any Lagrangian spheres. It immediately follows that the only smooth varieties X that support $C \times \mathbb{C}P^n$ as their hyperplane section must have small dual. For $n \leq 5$ results of Ein [17, 18] make it even possible to list all such X's.

We conclude with a remark on the methods. The symplectic approach outlined above gives coarser results. Indeed Sommese [35] provides examples of varieties that cannot be ample divisors whereas the methods above only rule out the possibility of being very ample. On the other hand the symplectic approach has an advantage in its robustness with respect to small deformations (see [9], c.f. [36]).

5.2. Degenerations of algebraic varieties

The methods of the previous section can also be used to study degenerations of algebraic varieties. Let Y be a smooth projective variety. We say that Y admits a Kähler degeneration with isolated singularities if there exists a Kähler manifold X and a proper holomorphic map $\pi : X \to D$ to the unit disc $D \subset \mathbb{C}$ with the following properties:

1. Every $0 \neq t \in D$ is a regular value of π (hence, all the fibres $X_t = \pi^{-1}(t)$, $t \neq 0$, are smooth Kähler manifolds).
2. 0 is a critical value of π and all the critical points of π are isolated.
3. Y is isomorphic (as a complex manifold) to one of the smooth fibres of π, say X_{t_0}, $t_0 \neq 0$.

As in the previous section this situation is related to symplectic geometry through the Lagrangian vanishing cycle construction. As pointed out by Seidel [39] one can locally morsify each of the critical points in $X_0 = \pi^{-1}(0)$ and then by applying Moser's argument obtain for each critical point of π at least one Lagrangian sphere in the nearby fibre X_ϵ. Since all the smooth fibres are symplectomorphic we obtain Lagrangian spheres also in Y.

Applying results from Section 4 to this situation we obtain examples of projective varieties that do not admit *any* degeneration with isolated singularities. For example, let Y be any of the following:

- $\mathbb{C}P^n$, $n \geq 2$. Or more generally $\mathbb{C}P^n \times M$, where M is a smooth variety with $\pi_2(M) = 0$ and $\dim_{\mathbb{C}} M \not\equiv n+1 \pmod{n+1}$.
- Any variety whose universal cover is $\mathbb{C}^n$, $(n \geq 2)$, or a domain in $\mathbb{C}^n$.

Then by the results in Section 4, Y has no Lagrangian spheres, hence does not admit any degeneration as above. More examples can be found in [9].

This point of view seems non-trivial especially when $H_n(Y;\mathbb{Z}) = 0$, where $n = \dim_{\mathbb{C}} Y$. In these cases the vanishing cycles are zero in homology and it seems that there are no obvious topological obstructions for degenerating Y as above. ¿From the list above, the first non-trivial example should be $\mathbb{C}P^n$ with $n = \text{odd} \geq 3$. It would be interesting to figure out to which extent the above statement could be

proved within the tools of pure algebraic geometry. Note that Lagrangian spheres are a non-algebraic object and it seems that their existence/non-existence cannot be formalized in purely algebro-geometric terms.

Another direction of applications should be to find an upper bound on the number of singular points of an algebraic variety X_0 that can be obtained from a degeneration of Y. Note that the vanishing cycles of different singular points of X_0 are disjoint. Thus the idea here is to obtain an upper bound on the number of possible disjoint Lagrangian spheres that can be embedded in Y. The simplest test case here should be the quadric $Q = \{z_0^2 + \cdots + z_{n+1}^2 = 0\} \subset \mathbb{C}P^{n+1}$, where $n \geq 2$. Clearly Q can be degenerated to a variety X_0 with isolated singularities (e.g. to a cone over a smaller dimensional quadric). It seems reasonable to expect that in every such degeneration the singular fibre X_0 will have only one singular point. Note that for $n =$ even this easily follows from topological reason but it may not be so when $n =$ odd ≥ 3 because $H_n(Q;\mathbb{Z}) = 0$. From a symplectic point of view the above statement would follow if we could prove that every two Lagrangian spheres in Q must intersect. This is currently still unknown but there are evidences supporting this conjecture [8]. It is likely that a refinement of the methods from [40] would be useful for this purpose. More generally, one could try to bound the number of singular fibres in a degeneration of other hypersurfaces $\Sigma \subset \mathbb{C}P^{n+1}$ (in terms of $\deg(\Sigma)$ and n). See [8, 9] for the conjectured bounds.

Acknowledgments. I am indebted to Leonid Polterovich for sharing with me his insight into symplectic intersections in general and especially regarding the problems presented in Section 3. Numerous discussions with him have influenced my conceptions on the subject. I wish to thank him also for valuable comments on earlier drafts of the paper. I would like to thank also Jonathan Wahl who told me about Sommese's work [35], Olivier Debarre for the reference to the works of Zak [43] and Ein [17, 18], and Paul Seidel for useful discussions on symplectic aspects of singularity theory and the material presented in section 5.

References

[1] M. Audin, *Fibrés normaux d'immersions en dimension double, points doubles d'immersions lagragiennes et plongements totalement réels.* Comment. Math. Helv. 63 (1988), 593–623.

[2] M. Audin, F. Lalonde & L. Polterovich, *Symplectic rigidity: Lagrangian submanifolds.* Holomorphic curves in symplectic geometry. Edited by M. Audin & J. Lafontaine. Progr. Math., 117. Birkhäuser 1994.

[3] V. Arnold, *The first steps of symplectic topology.* Uspekhi Mat. Nauk 41 (1986), no. 6 (252), 3–18, 229.

[4] V. Arnold, *Some remarks on symplectic monodromy of Milnor fibrations.* The Floer memorial volume, 99–103, Progr. Math., 133, Birkhäuser 1995.

[5] P. Biran, *Symplectic packing in dimension* 4, Geom. Funct. Anal. 7 (1997), 420–437.

[6] P. Biran, *A stability property of symplectic packing*, Invent. Math. 136 (1999), 123–155.
[7] P. Biran *Lagrangian barriers and symplectic embeddings.* Geom. Funct. Anal. 11 (2001), 407–464.
[8] P. Biran, *Lagrangian non-intersections.*, in preparation.
[9] P. Biran, *Symplectic obstructions in algebraic geometry.* in preparation.
[10] P. Biran & K. Cieliebak, *Symplectic topology on subcritical manifolds.* Comment. Math. Helv. 76 (2001), 712–753.
[11] P. Biran & K. Cieliebak, *Lagrangian embeddings into subcritical Stein manifolds.* Israel J. Math. 127 (2002), 221–244.
[12] M. Chaperon, *Quelques questions de géométrie symplectique.* Bourbaki seminar, Astérisque, 105–6, 231–249, Soc. Math. France, Paris, 1983.
[13] Y. Chekanov, *Critical points of quasifunctions, and generating families of Legendrian manifolds.* Funct. Anal. Appl. 30 (1996), 118–128.
[14] K. Cieliebak, *Subcritical Stein manifolds are split.* Preprint.
[15] S. Donaldson, *Polynomials, vanishing cycles and Floer homology.* Mathematics: frontiers and perspectives, 55–64, Amer. Math. Soc., 2000.
[16] A. Floer, *Morse theory for Lagrangian intersections.* J. Differential Geom. 28 (1988), 513–547.
[17] L. Ein, *Varieties with small dual varieties. I.* Invent. Math. 86 (1986), 63–74.
[18] L. Ein, *Varieties with small dual varieties. II.* Duke Math. J. 52 (1985), 895–907.
[19] Y. Eliashberg, A. Givental & H. Hofer *Introduction to symplectic field theory.* GAFA 2000 (Tel Aviv, 1999). Geom. Funct. Anal. 2000, Special Volume, Part II, 560–673.
[20] Y. Eliashberg & M. Gromov, *Lagrangian intersection theory: finite-dimensional approach.* Geometry of differential equations, 27–118, Amer. Math. Soc. Transl. Ser. 2, 186, Amer. Math. Soc. 1998.
[21] K. Fukaya, Y.-G. Oh, H. Ohta & K. Ono, *Lagrangian intersection Floer theory - anomaly and obstruction.* Preprint.
[22] M. Gromov, *Pseudoholomorphic curves in symplectic manifolds.* Invent. Math. 82 (1985), 307–347.
[23] H. Hofer, *Lagrangian embeddings and critical point theory.* Ann. Inst. H. Poincaré Anal. Non Linéaire 2 (1985), 407–462.
[24] S. Kleiman, *About the conormal scheme.* Complete intersections (Acireale, 1983), 161–197, Lecture Notes in Math., 1092, Springer 1984.
[25] F. Lalonde & J.-C Sikorav, *Sous-variétés lagrangiennes et lagrangiennes exactes des fibrés cotangents.* Comment. Math. Helv. 66 (1991), 18–33.
[26] F. Laudenbach & J.-C Sikorav, *Persistance d'intersection avec la section nulle au cours d'une isotopie hamiltonienne dans un fibré cotangent.* Invent. Math. 82 (1985), 349–357.
[27] D. McDuff & L. Polterovich, *Symplectic packings and algebraic geometry.* Invent. Math. 115 (1994), 405–434.
[28] D. McDuff and D. Salamon, *Introduction to symplectic topology.* Oxford Mathematical Monographs. Oxford University Press, New York, 1998.

[29] Y.-G. Oh, *Floer cohomology, spectral sequences, and the Maslov class of Lagrangian embeddings.* Internat. Math. Res. Notices 1996, 305–346.
[30] Y.-G. Oh, *Floer cohomology of Lagrangian intersections and pseudo-holomorphic disks. I.* Comm. Pure Appl.
[31] L. Polterovich, *Monotone Lagrange submanifolds of linear spaces and the Maslov class in cotangent bundles.* Math. Z. 207 (1991), 217–222.
[32] L. Polterovich, *Hofer's diameter and Lagrangian intersections.* Internat. Math. Res. Notices 1998, 217–223.
[33] L. Polterovich & K. F. Siburg, *Lagrangian submanifolds in convex domains of cotangent bundles.* Preprint.
[34] A. Silva, *Relative vanishing theorems. I. Applications to ample divisors.* Comment. Math. Helv. 52 (1977), 483–489.
[35] A. Sommese, *On manifolds that cannot be ample divisors.* Math. Ann. 221 (1976), 55–72.
[36] A. Sommese, *Nonsmoothable varieties.* Comment. Math. Helv. 54 (1979), no. 1, 140–146.
[37] D. Sullivan, *Cycles for the dynamical study of foliated manifolds and complex manifolds.* Invent. Math. 36 (1976), 225–255.
[38] P. Seidel, *Floer homology and the symplectic isotopy problem*, PhD thesis, Oxford University 1997.
[39] P. Seidel, *Graded Lagrangian submanifolds.* Bull. Soc. Math. France 128 (2000), 103–149.
[40] P. Seidel, *A long exact sequence for symplectic Floer cohomology.* Preprint.
[41] C. Viterbo, *A new obstruction to embedding Lagrangian tori.* Invent. Math. 100 (1990), 301–320.
[42] C. Viterbo, *Exact Lagrange submanifolds, periodic orbits and the cohomology of free loop spaces.* J. Differential Geom. 47 (1997), 420–468.
[43] F. Zak, *Tangents and secants of algebraic varieties.* Mathematical Monographs, 127. American Mathematical Society, 1993.

ICM 2002 · Vol. II · 257–271

Black Holes and the Penrose Inequality in General Relativity

Hubert L. Bray *

Abstract

In a paper [23] in 1973, R. Penrose made a physical argument that the total mass of a spacetime which contains black holes with event horizons of total area A should be at least $\sqrt{A/16\pi}$. An important special case of this physical statement translates into a very beautiful mathematical inequality in Riemannian geometry known as the Riemannian Penrose inequality. One particularly geometric aspect of this problem is the fact that apparent horizons of black holes in this setting correspond to minimal surfaces in Riemannian 3-manifolds. The Riemannian Penrose inequality was first proved by G. Huisken and T. Ilmanen in 1997 for a single black hole [17] and then by the author in 1999 for any number of black holes [6]. The two approaches use two different geometric flow techniques. The most general version of the Penrose inequality is still open.

In this talk we will sketch the author's proof by flowing Riemannian manifolds inside the class of asymptotically flat 3-manifolds (asymptotic to $\mathbf{R}^3$ at infinity) which have nonnegative scalar curvature and contain minimal spheres. This new flow of metrics has very special properties and simulates an initial physical situation in which all of the matter falls into the black holes which merge into a single, spherically symmetric black hole given by the Schwarzschild metric. Since the Schwarzschild metric gives equality in the Penrose inequality and the flow decreases the total mass while preserving the area of the horizons of the black holes, the Penrose inequality follows. We will also discuss how these techniques can be generalized in higher dimensions.

2000 Mathematics Subject Classification: 53, 83.
Keywords and Phrases: Black holes, Penrose inequality, Positive mass theorem, Quasi-local mass, General relativity.

1. Introduction

A natural interpretation of the Penrose inequality is that the mass contributed by a collection of black holes is (at least) $\sqrt{A/16\pi}$, where A is the total area of the event horizons of the black holes. More generally, the question "How much matter

*Mathematics Department, 2-179, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. E-mail: bray@math.mit.edu

is in a given region of a spacetime?" is still very much an open problem [12]. In this paper, we will discuss some of the qualitative aspects of mass in general relativity, look at examples which are informative, and sketch a proof of the Riemannian Penrose inequality.

1.1. Total mass in general relativity

Two notions of mass which are well understood in general relativity are local energy density at a point and the total mass of an asymptotically flat spacetime. However, defining the mass of a region larger than a point but smaller than the entire universe is not very well understood at all.

Suppose (M^3, g) is a Riemannian 3-manifold isometrically embedded in a (3+1) dimensional Lorentzian spacetime. Suppose that M^3 has zero second fundamental form in the spacetime. This is a simplifying assumption which allows us to think of (M^3, g) as a "$t = 0$" slice of the spacetime. The Penrose inequality (which allows for M^3 to have general second fundamental form) is known as the Riemannian Penrose inequality when the second fundamental form is set to zero.

We also want to only consider (M^3, g) that are asymptotically flat at infinity, which means that for some compact set K, the "end" $M^3 \backslash K$ is diffeomorphic to $\mathbf{R}^3 \backslash B_1(0)$, where the metric g is asymptotically approaching (with certain decay conditions) the standard flat metric δ_{ij} on $\mathbf{R}^3$ at infinity. The simplest example of an asymptotically flat manifold is $(\mathbf{R}^3, \delta_{ij})$ itself. Other good examples are the conformal metrics $(\mathbf{R}^3, u(x)^4 \delta_{ij})$, where $u(x)$ approaches a constant sufficiently rapidly at infinity. (Also, sometimes it is convenient to allow (M^3, g) to have multiple asymptotically flat ends, in which case each connected component of $M^3 \backslash K$ must have the property described above.)

The purpose of these assumptions on the asymptotic behavior of (M^3, g) at infinity is that they imply the existence of the limit

$$m = \frac{1}{16\pi} \lim_{\sigma \to \infty} \int_{S_\sigma} \sum_{i,j} (g_{ij,i} \nu_j - g_{ii,j} \nu_j) \, d\mu, \tag{1}$$

where S_σ is the coordinate sphere of radius σ, ν is the unit normal to S_σ, and $d\mu$ is the area element of S_σ in the coordinate chart. The quantity m is called the **total mass** (or ADM mass) of (M^3, g) (see [1], [2], [24], and [27]).

Instead of thinking of total mass as given by equation 1, it is better to consider the following example. Going back to the example $(\mathbf{R}^3, u(x)^4 \delta_{ij})$, if we suppose that $u(x) > 0$ has the asymptotics at infinity

$$u(x) = a + b/|x| + \mathcal{O}(1/|x|^2) \tag{2}$$

(and derivatives of the $\mathcal{O}(1/|x|^2)$ term are $\mathcal{O}(1/|x|^3)$), then the total mass of (M^3, g) is

$$m = 2ab. \tag{3}$$

Furthermore, suppose (M^3, g) is any metric whose "end" is isometric to $(\mathbf{R}^3 \backslash K, u(x)^4 \delta_{ij})$, where $u(x)$ is harmonic in the coordinate chart of the end $(\mathbf{R}^3 \backslash K, \delta_{ij})$

and goes to a constant at infinity. Then expanding $u(x)$ in terms of spherical harmonics demonstrates that $u(x)$ satisfies condition 2. We will call these Riemannian manifolds (M^3, g) **harmonically flat at infinity**, and we note that the total mass of these manifolds is also given by equation 3.

A very nice lemma by Schoen and Yau is that, given any $\epsilon > 0$, it is always possible to perturb an asymptotically flat manifold to become harmonically flat at infinity such that the total mass changes less than ϵ and the metric changes less than ϵ pointwise, all while maintaining nonnegative scalar curvature (discussed in a moment). Hence, it happens that to prove the theorems in this paper, we only need to consider harmonically flat manifolds! Thus, we can use equation 3 as our definition of total mass. As an example, note that $(\mathbf{R}^3, \delta_{ij})$ has zero total mass. Also, note that, qualitatively, the total mass of an asymptotically flat or harmonically flat manifold is the $1/r$ rate at which the metric becomes flat at infinity.

1.2. Local energy density

Another quantification of mass which is well understood is local energy density. In fact, in this setting, the local energy density at each point is

$$\mu = \frac{1}{16\pi} R, \tag{4}$$

where R is the scalar curvature of the 3-manifold (which has zero second fundamental form in the spacetime) at each point. Thus, we note that $(\mathbf{R}^3, \delta_{ij})$ has zero energy density at each point as well as zero total mass. This is appropriate since $(\mathbf{R}^3, \delta_{ij})$ is in fact a "$t = 0$" slice of Minkowski spacetime, which represents a vacuum. Classically, physicists consider $\mu \geq 0$ to be a physical assumption. Hence, from this point on, we will not only assume that (M^3, g) is asymptotically flat, but also that it has nonnegative scalar curvature,

$$R \geq 0. \tag{5}$$

This notion of energy density also helps us understand total mass better. After all, we can take any asymptotically flat manifold and then change the metric to be perfectly flat outside a large compact set, thereby giving the new metric zero total mass. However, if we introduce the physical condition that both metrics have nonnegative scalar curvature, then it is a beautiful theorem that this is in fact not possible, unless the original metric was already $(\mathbf{R}^3, \delta_{ij})$! (This theorem is actually a corollary to the positive mass theorem discussed in a moment.) Thus, the curvature obstruction of having nonnegative scalar curvature at each point is a very interesting condition.

Also, notice the indirect connection between the total mass and local energy density. At this point, there does not seem to be much of a connection at all. Total mass is the $1/r$ rate at which the metric becomes flat at infinity, and local energy density is the scalar curvature at each point. Furthermore, if a metric is changed in a compact set, local energy density is changed, but the total mass is unaffected.

The reason for this is that the total mass is *not* the integral of the local energy density over the manifold. In fact, this integral fails to take potential energy into account (which would be expected to contribute a negative energy) as well as gravitational energy (discussed in a moment). Hence, it is not initially clear what we should expect the relationship between total mass and local energy density to be, so let us begin with an example.

1.3. Example using superharmonic functions in $\mathbf{R}^3$

Once again, let us return to the $(\mathbf{R}^3, u(x)^4\delta_{ij})$ example. The formula for the scalar curvature is

$$R = -8u(x)^{-5}\Delta u(x). \tag{6}$$

Hence, since the physical assumption of nonnegative energy density implies nonnegative scalar curvature, we see that $u(x) > 0$ must be superharmonic ($\Delta u \leq 0$). For simplicity, let's also assume that $u(x)$ is harmonic outside a bounded set so that we can expand $u(x)$ at infinity using spherical harmonics. Hence, $u(x)$ has the asymptotics of equation 2. By the maximum principle, it follows that the minimum value for $u(x)$ must be a, referring to equation 2. Hence, $b \geq 0$, which implies that $m \geq 0$! Thus we see that the assumption of nonnegative energy density at each point of $(\mathbf{R}^3, u(x)^4\delta_{ij})$ implies that the total mass is also nonnegative, which is what one would hope.

1.4. The positive mass theorem

More generally, suppose we have any asymptotically flat manifold with nonnegative scalar curvature, is it true that the total mass is also nonnegative? The answer is *yes*, and this fact is know as the positive mass theorem, first proved by Schoen and Yau [25] in 1979 using minimal surface techniques and then by Witten [30] in 1981 using spinors.

Theorem 1 *(Schoen-Yau) Let (M^3, g) be any asymptotically flat, complete Riemannian manifold with nonnegative scalar curvature. Then the total mass $m \geq 0$, with equality if and only if (M^3, g) is isometric to $(\mathbf{R}^3, \delta)$.*

1.5. Black holes

Another very interesting and natural phenomenon in general relativity is the existence of black holes. Instead of thinking of black holes as singularities in a spacetime, we will think of black holes in terms of their horizons. Given a surface in a spacetime, suppose that it admits an outward shell of light. If the surface area of this shell of light is decreasing everywhere on the surface, then this is called a trapped surface. The outermost boundary of these trapped surfaces is called the apparent horizon of the black hole. Apparent horizons can be computed based on their local geometry, and an apparent horizon always implies the existence of an event horizon outside of it [15].

Now let us return to the case we are considering in this paper where (M^3, g) is a "t = 0" slice of a spacetime with zero second fundamental form. Then it is a

very nice geometric fact that apparent horizons of black holes intersected with M^3 correspond to the connected components of the outermost minimal surface Σ_0 of (M^3, g).

All of the surfaces we are considering in this paper will be required to be smooth boundaries of open bounded regions, so that outermost is well-defined with respect to a chosen end of the manifold [6]. A minimal surface in (M^3, g) is a surface which is a critical point of the area function with respect to any smooth variation of the surface. The first variational calculation implies that minimal surfaces have zero mean curvature. The surface Σ_0 of (M^3, g) is defined as the boundary of the union of the open regions bounded by all of the minimal surfaces in (M^3, g). It turns out that Σ_0 also has to be a minimal surface, so we call Σ_0 the **outermost minimal surface**.

We will also define a surface to be **(strictly) outer minimizing** if every surface which encloses it has (strictly) greater area. Note that outermost minimal surfaces are strictly outer minimizing. Also, we define a **horizon** in our context to be any minimal surface which is the boundary of a bounded open region.

It also follows from a stability argument (using the Gauss-Bonnet theorem interestingly) that each component of a stable minimal surface (in a 3-manifold with nonnegative scalar curvature) must have the topology of a sphere. Furthermore, there is a physical argument, based on [23], which suggests that the mass contributed by the black holes (thought of as the connected components of Σ_0) should be defined to be $\sqrt{A_0/16\pi}$, where A_0 is the area of Σ_0. Hence, the physical argument that the total mass should be greater than or equal to the mass contributed by the black holes yields that following geometric statement.

The Riemannian Penrose Inequality
Let (M^3, g) be a complete, smooth, 3-manifold with nonnegative scalar curvature which is harmonically flat at infinity with total mass m and which has an outermost minimal surface Σ_0 of area A_0. Then

$$m \geq \sqrt{\frac{A_0}{16\pi}}, \tag{7}$$

with equality if and only if (M^3, g) is isometric to the Schwarzschild metric $(\mathbf{R}^3 \backslash \{0\}, (1 + \frac{m}{2|x|})^4 \delta_{ij})$ outside their respective outermost minimal surfaces.

The above statement has been proved by the author [6], and by Huisken and Ilmanen [17] where A_0 is defined instead to be the area of the largest connected component of Σ_0. We will discuss both approaches in this paper, which are very different, although they both involve flowing surfaces and/or metrics.

We also clarify that the above statement is with respect to a chosen end of (M^3, g), since both the total mass and the definition of outermost refer to a particular end. In fact, nothing very important is gained by considering manifolds with more than one end, since extra ends can always be compactified by connect summing them (around a neighborhood of infinity) with large spheres while still preserving nonnegative scalar curvature, for example. Hence, we will typically consider manifolds with just one end. In the case that the manifold has multiple ends,

we will require every surface (which could have multiple connected components) in this paper to enclose all of the ends of the manifold except the chosen end.

Other contributions on the Penrose Conjecture have also been made by Herzlich [16] using the Dirac operator which Witten [30] used to prove the positive mass theorem, by Gibbons [14] in the special case of collapsing shells, by Tod [29], by Bartnik [4] for quasi-spherical metrics, and by the author [7] using isoperimetric surfaces. There is also some interesting work of Ludvigsen and Vickers [21] using spinors and Bergqvist [5], both concerning the Penrose inequality for null slices of a space-time.

1.6. The Schwarzschild metric

The Schwarzschild metric $(\mathbf{R}^3\backslash\{0\}, (1+\frac{m}{2|x|})^4\delta_{ij})$, referred to in the above statement of the Riemannian Penrose Inequality, is a particularly important example to consider, and corresponds to a zero-second fundamental form, space-like slice of the usual (3+1)-dimensional Schwarzschild metric (which represents a spherically symmetric static black hole in vacuum). The 3-dimensional Schwarzschild metrics have total mass $m > 0$ and are characterized by being the only spherically symmetric, geodesically complete, zero scalar curvature 3-metrics, other than $(\mathbf{R}^3, \delta_{ij})$. They can also be embedded in 4-dimensional Euclidean space (x, y, z, w) as the set of points satisfying $|(x, y, z)| = \frac{w^2}{8m} + 2m$, which is a parabola rotated around an S^2. This last picture allows us to see that the Schwarzschild metric, which has two ends, has a Z_2 symmetry which fixes the sphere with $w = 0$ and $|(x, y, z)| = 2m$, which is clearly minimal. Furthermore, the area of this sphere is $4\pi(2m)^2$, giving equality in the Riemannian Penrose Inequality.

2. The conformal flow of metrics

Given any initial Riemannian manifold (M^3, g_0) which has nonnegative scalar curvature and which is harmonically flat at infinity, we will define a continuous, one parameter family of metrics (M^3, g_t), $0 \leq t < \infty$. This family of metrics will converge to a 3-dimensional Schwarzschild metric and will have other special properties which will allow us to prove the Riemannian Penrose Inequality for the original metric (M^3, g_0).

In particular, let Σ_0 be the outermost minimal surface of (M^3, g_0) with area A_0. Then we will also define a family of surfaces $\Sigma(t)$ with $\Sigma(0) = \Sigma_0$ such that $\Sigma(t)$ is minimal in (M^3, g_t). This is natural since as the metric g_t changes, we expect that the location of the horizon $\Sigma(t)$ will also change. Then the interesting quantities to keep track of in this flow are $A(t)$, the total area of the horizon $\Sigma(t)$ in (M^3, g_t), and $m(t)$, the total mass of (M^3, g_t) in the chosen end.

In addition to all of the metrics g_t having nonnegative scalar curvature, we will also have the very nice properties that

$$A'(t) = 0, \tag{8}$$

$$m'(t) \leq 0 \tag{9}$$

for all $t \geq 0$. Then since (M^3, g_t) converges to a Schwarzschild metric (in an appropriate sense) which gives equality in the Riemannian Penrose Inequality as described in the introduction,

$$m(0) \geq m(\infty) = \sqrt{\frac{A(\infty)}{16\pi}} = \sqrt{\frac{A(0)}{16\pi}} \tag{10}$$

which proves the Riemannian Penrose Inequality for the original metric (M^3, g_0). The hard part, then, is to find a flow of metrics which preserves nonnegative scalar curvature and the area of the horizon, decreases total mass, and converges to a Schwarzschild metric as t goes to infinity.

2.1. The definition of the flow

In fact, the metrics g_t will all be conformal to g_0. This conformal flow of metrics can be thought of as the solution to a first order o.d.e. in t defined by equations 11, 12, 13, and 14. Let

$$g_t = u_t(x)^4 g_0 \tag{11}$$

and $u_0(x) \equiv 1$. Given the metric g_t, define

$$\Sigma(t) = \text{the outermost minimal area enclosure of } \Sigma_0 \text{ in } (M^3, g_t) \tag{12}$$

where Σ_0 is the original outer minimizing horizon in (M^3, g_0). In the cases in which we are interested, $\Sigma(t)$ will not touch Σ_0, from which it follows that $\Sigma(t)$ is actually a strictly outer minimizing horizon of (M^3, g_t). Then given the horizon $\Sigma(t)$, define $v_t(x)$ such that

$$\begin{cases} \Delta_{g_0} v_t(x) \equiv 0 & \text{outside } \Sigma(t) \\ v_t(x) = 0 & \text{on } \Sigma(t) \\ \lim_{x \to \infty} v_t(x) = -e^{-t} & \end{cases} \tag{13}$$

and $v_t(x) \equiv 0$ inside $\Sigma(t)$. Finally, given $v_t(x)$, define

$$u_t(x) = 1 + \int_0^t v_s(x) ds \tag{14}$$

so that $u_t(x)$ is continuous in t and has $u_0(x) \equiv 1$.

Note that equation 14 implies that the first order rate of change of $u_t(x)$ is given by $v_t(x)$. Hence, the first order rate of change of g_t is a function of itself, g_0, and $v_t(x)$ which is a function of g_0, t, and $\Sigma(t)$ which is in turn a function of g_t and Σ_0. Thus, the first order rate of change of g_t is a function of t, g_t, g_0, and Σ_0.

Theorem 2 *Taken together, equations 11, 12, 13, and 14 define a first order o.d.e. in t for $u_t(x)$ which has a solution which is Lipschitz in the t variable, C^1 in the x variable everywhere, and smooth in the x variable outside $\Sigma(t)$. Furthermore, $\Sigma(t)$ is a smooth, strictly outer minimizing horizon in (M^3, g_t) for all $t \geq 0$, and $\Sigma(t_2)$ encloses but does not touch $\Sigma(t_1)$ for all $t_2 > t_1 \geq 0$.*

Since $v_t(x)$ is a superharmonic function in (M^3, g_0) (harmonic everywhere except on $\Sigma(t)$, where it is weakly superharmonic), it follows that $u_t(x)$ is superharmonic as well. Thus, from equation 14 we see that $\lim_{x\to\infty} u_t(x) = e^{-t}$ and consequently that $u_t(x) > 0$ for all t by the maximum principle. Then since

$$R(g_t) = u_t(x)^{-5}(-8\Delta_{g_0} + R(g_0))u_t(x), \tag{15}$$

it follows that (M^3, g_t) is an asymptotically flat manifold with nonnegative scalar curvature.

Even so, it still may not seem like g_t is particularly naturally defined since the rate of change of g_t appears to depend on t and the original metric g_0 in equation 13. We would prefer a flow where the rate of change of g_t can be defined purely as a function of g_t (and Σ_0 perhaps), and interestingly enough this actually does turn out to be the case. In section 2.4. we prove this very important fact and define a new equivalence class of metrics called the harmonic conformal class. Then once we decide to find a flow of metrics which stays inside the harmonic conformal class of the original metric (outside the horizon) and keeps the area of the horizon $\Sigma(t)$ constant, then we are basically forced to choose the particular conformal flow of metrics defined above.

Theorem 3 *The function $A(t)$ is constant in t and $m(t)$ is non-increasing in t, for all $t \geq 0$.*

The fact that $A'(t) = 0$ follows from the fact that to first order the metric is not changing on $\Sigma(t)$ (since $v_t(x) = 0$ there) and from the fact that to first order the area of $\Sigma(t)$ does not change as it moves outward since $\Sigma(t)$ is a critical point for area in (M^3, g_t). Hence, the interesting part of theorem 3 is proving that $m'(t) \leq 0$. Curiously, this follows from a nice trick using the Riemannian positive mass theorem, which we describe in section 2.3..

Another important aspect of this conformal flow of the metric is that outside the horizon $\Sigma(t)$, the manifold (M^3, g_t) becomes more and more spherically symmetric and "approaches" a Schwarzschild manifold $(\mathbf{R}^3\backslash\{0\}, s)$ in the limit as t goes to ∞. More precisely,

Theorem 4 *For sufficiently large t, there exists a diffeomorphism ϕ_t between (M^3, g_t) outside the horizon $\Sigma(t)$ and a fixed Schwarzschild manifold $(\mathbf{R}^3\backslash\{0\}, s)$ outside its horizon. Furthermore, for all $\epsilon > 0$, there exists a T such that for all $t > T$, the metrics g_t and $\phi_t^*(s)$ (when determining the lengths of unit vectors of (M^3, g_t)) are within ϵ of each other and the total masses of the two manifolds are within ϵ of each other. Hence,*

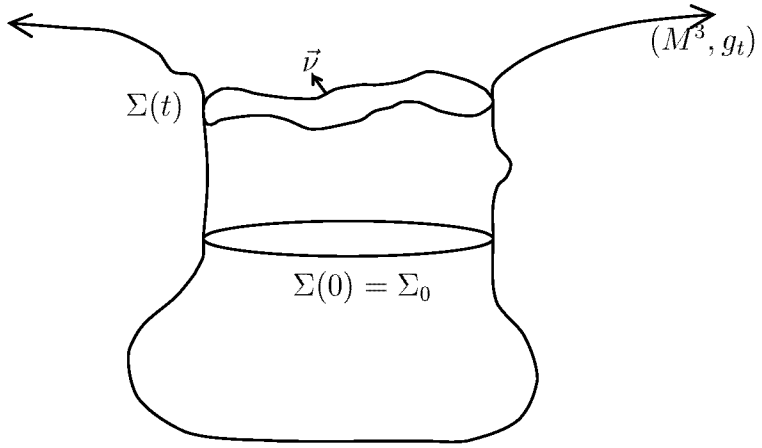
$$\lim_{t\to\infty} \frac{m(t)}{\sqrt{A(t)}} = \sqrt{\frac{1}{16\pi}}. \tag{16}$$

Theorem 4 is not that surprising really although a careful proof is reasonably long. However, if one is willing to believe that the flow of metrics converges to a spherically symmetric metric outside the horizon, then theorem 4 follows from two

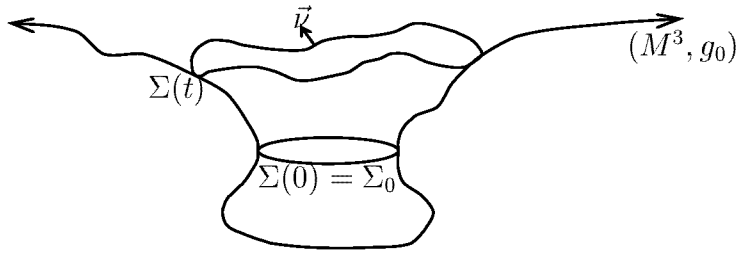
facts. The first fact is that the scalar curvature of (M^3, g_t) eventually becomes identically zero outside the horizon $\Sigma(t)$ (assuming (M^3, g_0) is harmonically flat). This follows from the facts that $\Sigma(t)$ encloses any compact set in a finite amount of time, that harmonically flat manifolds have zero scalar curvature outside a compact set, that $u_t(x)$ is harmonic outside $\Sigma(t)$, and equation 15. The second fact is that the Schwarzschild metrics are the only complete, spherically symmetric 3-manifolds with zero scalar curvature (except for the flat metric on R^3).

The Riemannian Penrose inequality, inequality 7, then follows from equation 10 using theorems 2, 3 and 4, for harmonically flat manifolds [6]. Since asymptotically flat manifolds can be approximated arbitrarily well by harmonically flat manifolds while changing the relevant quantities arbitrarily little, the asymptotically flat case also follows. Finally, the case of equality of the Penrose inequality follows from a more careful analysis of these same arguments.

2.2. Qualitative discussion



The diagrams above and below are meant to help illustrate some of the properties of the conformal flow of the metric. The above picture is the original metric which has a strictly outer minimizing horizon Σ_0. As t increases, $\Sigma(t)$ moves outwards, but never inwards. In the diagram below, we can observe one of the consequences of the fact that $A(t) = A_0$ is constant in t. Since the metric is not changing inside $\Sigma(t)$, all of the horizons $\Sigma(s)$, $0 \leq s \leq t$ have area A_0 in (M^3, g_t). Hence, inside $\Sigma(t)$, the manifold (M^3, g_t) becomes cylinder-like in the sense that it is laminated (meaning foliated but with some gaps allowed) by all of the previous horizons which all have the same area A_0 with respect to the metric g_t.



Now let us suppose that the original horizon Σ_0 of (M^3, g) had two components, for example. Then each of the components of the horizon will move outwards as t increases, and at some point before they touch they will suddenly jump outwards to form a horizon with a single component enclosing the previous horizon with two components. Even horizons with only one component will sometimes jump outwards, but no more than a countable number of times. It is interesting that this phenomenon of surfaces jumping is also found in the Huisken-Ilmanen approach to the Penrose conjecture using their generalized $1/H$ flow.

2.3. Proof that $m'(t) \leq 0$

The most surprising aspect of the flow defined in section 2.1. is that $m'(t) \leq 0$. As mentioned in that section, this important fact follows from a nice trick using the Riemannian positive mass theorem.

The first step is to realize that while the rate of change of g_t appears to depend on t and g_0, this is in fact an illusion. As is described in detail in section 2.4., the rate of change of g_t can be described purely in terms of g_t (and Σ_0). It is also true that the rate of change of g_t depends only on g_t and $\Sigma(t)$. Hence, there is no special value of t, so proving $m'(t) \leq 0$ is equivalent to proving $m'(0) \leq 0$. Thus, without loss of generality, we take $t = 0$ for convenience.

Now expand the harmonic function $v_0(x)$, defined in equation 13, using spherical harmonics at infinity, to get

$$v_0(x) = -1 + \frac{c}{|x|} + \mathcal{O}\left(\frac{1}{|x|^2}\right) \tag{17}$$

for some constant c. Since the rate of change of the metric g_t at $t = 0$ is given by $v_0(x)$ and since the total mass $m(t)$ depends on the $1/r$ rate at which the metric g_t becomes flat at infinity (see equation 3), it is not surprising that direct calculation gives us that

$$m'(0) = 2(c - m(0)). \tag{18}$$

Hence, to show that $m'(0) \leq 0$, we need to show that

$$c \leq m(0). \tag{19}$$

In fact, counterexamples to equation 19 can be found if we remove either of the requirements that $\Sigma(0)$ (which is used in the definition of $v_0(x)$) be a minimal

surface or that (M^3, g_0) have nonnegative scalar curvature. Hence, we quickly see that equation 19 is a fairly deep conjecture which says something quite interesting about manifold with nonnegative scalar curvature. Well, the Riemannian positive mass theorem is also a deep conjecture which says something quite interesting about manifolds with nonnegative scalar curvature. Hence, it is natural to try to use the Riemannian positive mass theorem to prove equation 19.

Thus, we want to create a manifold whose total mass depends on c from equation 17. The idea is to use a reflection trick similar to one used by Bunting and Masood-ul-Alam for another purpose in [11]. First, remove the region of M^3 inside $\Sigma(0)$ and then reflect the remainder of (M^3, g_0) through $\Sigma(0)$. Define the resulting Riemannian manifold to be $(\bar{M}^3, \bar{g}_0)$ which has two asymptotically flat ends since (M^3, g_0) has exactly one asymptotically flat end not contained by $\Sigma(0)$. Note that $(\bar{M}^3, \bar{g}_0)$ has nonnegative scalar curvature everywhere except on $\Sigma(0)$ where the metric has corners. In fact, the fact that $\Sigma(0)$ has zero mean curvature (since it is a minimal surface) implies that $(\bar{M}^3, \bar{g}_0)$ has *distributional* nonnegative scalar curvature everywhere, even on $\Sigma(0)$. This notion is made rigorous in [6]. Thus we have used the fact that $\Sigma(0)$ is minimal in a critical way.

Recall from equation 13 that $v_0(x)$ was defined to be the harmonic function equal to zero on $\Sigma(0)$ which goes to -1 at infinity. We want to reflect $v_0(x)$ to be defined on all of $(\bar{M}^3, \bar{g}_0)$. The trick here is to define $v_0(x)$ on $(\bar{M}^3, \bar{g}_0)$ to be the harmonic function which goes to -1 at infinity in the original end and goes to 1 at infinity in the reflect end. By symmetry, $v_0(x)$ equals 0 on $\Sigma(0)$ and so agrees with its original definition on (M^3, g_0).

The next step is to compactify one end of $(\bar{M}^3, \bar{g}_0)$. By the maximum principle, we know that $v_0(x) > -1$ and $c > 0$, so the new Riemannian manifold $(\bar{M}^3, (v_0(x)+1)^4\bar{g}_0)$ does the job quite nicely and compactifies the original end to a point. In fact, the compactified point at infinity and the metric there can be filled in smoothly (using the fact that (M^3, g_0) is harmonically flat). It then follows from equation 15 that this new compactified manifold has nonnegative scalar curvature since $v_0(x)+1$ is harmonic.

The last step is simply to apply the Riemannian positive mass theorem to $(\bar{M}^3, (v_0(x)+1)^4\bar{g}_0)$. It is not surprising that the total mass $\tilde{m}(0)$ of this manifold involves c, but it is quite lucky that direct calculation yields

$$\tilde{m}(0) = -4(c - m(0)), \tag{20}$$

which must be positive by the Riemannian positive mass theorem. Thus, we have that

$$m'(0) = 2(c - m(0)) = -\frac{1}{2}\tilde{m}(0) \leq 0. \tag{21}$$

2.4. The harmonic conformal class of a metric

As a final topic which is also of independent interest, we define a new equivalence class and partial ordering of conformal metrics. These new objects provide a natural motivation for studying conformal flows of metrics to try to prove the

Riemannian Penrose inequality. Let

$$g_2 = u(x)^{\frac{4}{n-2}} g_1, \tag{22}$$

where g_2 and g_1 are metrics on an n-dimensional manifold M^n, $n \geq 3$. Then we get the surprisingly simple identity that

$$\Delta_{g_1}(u\phi) = u^{\frac{n+2}{n-2}} \Delta_{g_2}(\phi) + \phi \Delta_{g_1}(u) \tag{23}$$

for any smooth function ϕ. This motivates us to define the following relation.

Definition 1 *Define*

$$g_2 \sim g_1$$

if and only if equation 22 is satisfied with $\Delta_{g_1}(u) = 0$ *and* $u(x) > 0$.

Then from equation 23 we get the following lemma.

Lemma 1 *The relation* $\sim$ *is reflexive, symmetric, and transitive, and hence is an equivalence relation.*

Thus, we can define the following equivalence class of metrics.

Definition 2 *Define*

$$[g]_H = \{\bar{g} \mid \bar{g} \sim g\}$$

to be the **harmonic conformal class** *of the metric* g.

Of course, this definition is most interesting when (M^n, g) has nonconstant positive harmonic functions, which happens for example when (M^n, g) has a boundary.

Also, we can modify the relation $\sim$ to get another relation $\succeq$.

Definition 3 *Define*

$$g_2 \succeq g_1$$

if and only if equation 22 is satisfied with $-\Delta_{g_1}(u) \geq 0$ *and* $u(x) > 0$.

Then from equation 23 we get the following lemma.

Lemma 2 *The relation* $\succeq$ *is reflexive and transitive, and hence is a partial ordering.*

Since $\succeq$ is defined in terms of superharmonic functions, we will call it the superharmonic partial ordering of metrics on M^n. Then it is natural to define the following set of metrics.

Definition 4 *Define*

$$[g]_S = \{\bar{g} \mid \bar{g} \succeq g\}.$$

This set of metrics has the property that if $\bar{g} \in [g]_S$, then $[\bar{g}]_S \subset [g]_S$

Also, the scalar curvature transforms nicely under a conformal change of the metric. In fact, assuming equation 22 again,

$$R(g_2) = u(x)^{-(\frac{n+2}{n-2})} \left(-c_n \Delta_{g_1} + R(g_1)\right) u(x) \tag{24}$$

where $c_n = \frac{4(n-1)}{n-2}$. This gives us the following lemma.

Lemma 3 *The sign of the scalar curvature is preserved pointwise by* $\sim$. *That is, if* $g_2 \sim g_1$, *then* $sgn(R(g_2)(x)) = sgn(R(g_1)(x))$ *for all* $x \in M^n$. *Also, if* $g_2 \succeq g_1$, *and* g_1 *has non-negative scalar curvature, then* g_2 *has non-negative scalar curvature.*

Hence, the harmonic conformal equivalence relation $\sim$ and the superharmonic partial ordering $\succeq$ are useful for studying questions about scalar curvature. In particular, these notions are useful for studying the Riemannian Penrose inequality which concerns asymptotically flat 3-manifolds (M^3, g) with non-negative scalar curvature. Given such a manifold, define $m(g)$ to be the total mass of (M^3, g) and $A(g)$ to be the area of the outermost horizon (which could have multiple components) of (M^3, g). Define $P(g) = \frac{m(g)}{\sqrt{A(g)}}$ to be the Penrose quotient of (M^3, g). Then an interesting question is to ask which metric in $[g]_S$ minimizes $P(g)$.

Section 2. of this paper can be viewed as an answer to the above question. We showed that there exists a conformal flow of metrics (starting with g_0) for which the Penrose quotient was non-increasing, and in fact this conformal flow stays inside $[g_0]_S$. Furthermore, $g_{t_2} \in [g_{t_1}]_S$ for all $t_2 \geq t_1 \geq 0$. We showed that no matter which metric we start with, the metric converges to a Schwarzschild metric outside its horizon. Hence, the minimum value of $P(g)$ in $[g]_S$ is achieved in the limit by metrics converging to a Schwarzschild metric (outside their respective horizons).

In the case that g is harmonically flat at infinity, a Schwarzschild metric (outside the horizon) is contained in $[g]_S$. More generally, given any asymptotically flat manifold (M^3, g), we can use $\mathbf{R}^3 \backslash B_r(0)$ as a coordinate chart for the asymptotically flat end of (M^3, g) which we are interested in, where the metric g_{ij} approaches δ_{ij} at infinity in this coordinate chart. Then we can consider the conformal metric

$$g_C = \left(1 + \frac{C}{|x|}\right)^4 g \tag{25}$$

in this end. In the limit as C goes to infinity, the horizon will approach the coordinate sphere of radius C. Then outside this horizon in the limit as C goes to infinity, the function $(1 + \frac{C}{|x|})$ will be close to a superharmonic function on (M^3, g) and the metric g_C will approach a Schwarzschild metric (since the metric g is approaching the standard metric on $\mathbf{R}^3$). Hence, the Penrose quotient of g_C will approach $(16\pi)^{-1/2}$, which is the Penrose quotient of a Schwarzschild metric.

As a final note, we prove that the first order o.d.e. for $\{g_t\}$ defined in equations 11, 12, 13, and 14 is naturally defined in the sense that the rate of change of g_t is a function only of g_t and not of g_0 or t. To see this, given any solution $g_t = u_t(x)^4 g_0$ to equations 11, 12, 13, and 14, choose any $s > 0$ and define $\bar{u}_t(x) = u_t(x)/u_s(x)$ so that

$$g_t = \bar{u}_t(x)^4 g_s \tag{26}$$

and $\bar{u}_s(x) \equiv 1$. Then define $\bar{v}_t(x)$ such that

$$\begin{cases} \Delta_{g_s} \bar{v}_t(x) \equiv 0 & \text{outside } \Sigma(t) \\ \bar{v}_t(x) = 0 & \text{on } \Sigma(t) \\ \lim_{x \to \infty} \bar{v}_t(x) = -e^{-(t-s)} & \end{cases} \tag{27}$$

and $\bar{v}_t(x) \equiv 0$ inside $\Sigma(t)$. Then what we want to show is

$$\bar{u}_t(x) = 1 + \int_s^t \bar{v}_r(x) dr \tag{28}$$

To prove the above equation, we observe that from equations 23, 27, and 13 it follows that

$$v_t(x) = \bar{v}_t(x)\, u_s(x) \tag{29}$$

since $\lim_{x\to\infty} u_s(x) = e^{-s}$. Hence, since

$$u_t(x) = u_s(x) + \int_s^t v_r(x) dr \tag{30}$$

by equation 14, dividing through by $u_s(x)$ yields equation 28 as desired. Thus, we see that the rate of change of $g_t(x)$ at $t = s$ is a function of $\bar{v}_s(x)$ which in turn is just a function of $g_s(x)$ and the horizon $\Sigma(s)$. Hence, to understand properties of the flow we need only analyze the behavior of the flow for t close to zero, since any metric in the flow may be chosen to be the base metric.

References

[1] R. Arnowitt, S. Deser and C. Misner, Coordinate Invariance and Energy Expressions in General Relativity, *Phys. Rev.* **122** (1961), 997–1006.

[2] R. Bartnik, The Mass of an Asymptotically Flat Manifold, *Comm. Pure Appl. Math.* **39** (1986), 661–693.

[3] R. Bartnik, New Definition of Quasi-Local Mass, *Phys. Rev. Lett.* **62** (1989), 2346.

[4] R. Bartnik, Quasi-Spherical Metrics and Prescribed Scalar Curvature, *J. Diff. Geom.* **37** (1993), 31–71.

[5] G. Bergquist, On the Penrose Inequality and the Role of Auxiliary Spinor Fields, *Class. Quantum Grav.* **14** (1997), 2577–2583.

[6] H. L. Bray, Proof of the Riemannian Penrose Inequality Using the Positive Mass Theorem, *Jour. Diff. Geom.* (to appear).

[7] H. L. Bray, The Penrose Inequality in General Relativity and Volume Comparison Theorems Involving Scalar Curvature, thesis, Stanford University, 1997.

[8] H. L. Bray, F. Finster, Curvature Estimates and the Positive Mass Theorem, *Communications in Analysis and Geometry* (to appear).

[9] H. L. Bray, K. Iga, A Nonlinear Property of Superharmonic Functions in $\mathbf{R}^n$ and the Penrose Inequality in General Relativity, *Communications in Analysis and Geometry* (to appear).

[10] H. L. Bray, R. M. Schoen, Recent Proofs of the Riemannian Penrose Conjecture, *Current Developments in Mathematics 1999*, edited by S.-T. Yau.

[11] Bunting, Masood-ul-Alam, Non-Existence of Multiple Black Holes in Asymptotically Euclidean Static Vacuum Space-Time, *Gen. Rel. and Grav.*, Vol. 19, No. 2, 1987.

[12] D. Christodoulou and S.-T. Yau, Some Remarks on the Quasi-Local Mass, *Contemporary Mathematics* **71** (1988), 9–14.
[13] R. Geroch, Energy Extraction, *Ann. New York Acad. Sci.* **224** (1973), 108–17.
[14] G. Gibbons, Collapsing Shells and the Isoperimetric Inequality for Black Holes, *Class. Quant. Grav.* **14** (1997), 2905.
[15] S. W. Hawking and G. F. R. Ellis, *The Large-Scale Structure of Space-Time*, Cambridge University Press, Cambridge, 1973.
[16] M. Herzlich, A Penrose-like Inequality for the Mass of Riemannian Asymptotically Flat Manifolds, *Comm. Math. Phys.* **188** (1997), 121.
[17] G. Huisken and T. Ilmanen, The Inverse Mean Curvature Flow and the Riemannian Penrose Inequality, *J. Diff. Geom* (to appear).
[18] G. Huisken and T. Ilmanen, The Riemannian Penrose Inequality, *Int. Math. Res. Not.* **20** (1997), 1045–1058.
[19] G. Huisken and T. Ilmanen, A Note on Inverse Mean Curvature Flow, *Proceedings of the Workshop on Nonlinear Partial Differential Equations (Saitama University, Sept. 1997)*, available from Saitama University.
[20] P. S. Jang and R. M. Wald, The Positive Energy Conjecture and the Cosmic Censor Hypothesis, *J. Math. Phys.* **18** (1977), 41–44.
[21] M. Ludvigsen and J. Vickers, 1983, *J. Phys. A: Math. Gen.* **16**, 3349.
[22] T. H. Parker and C. H. Taubes, On Witten's Proof of the Positive Energy Theorem, *Commun. Math. Phys.* **84** (1982), 223–238.
[23] R. Penrose, Naked Singularities, *Ann. New York Acad. Sci.* **224** (1973), 125–134.
[24] R. Schoen, Variational Theory for the Total Scalar Curvature Functional for Riemannian Metrics and Related Topics,*Topics in Calculus of Variations (M. Giaquinta, ed.) Lecture Notes in Math.*, **1365**, 120-1-54, Springer, Berlin, 1987.
[25] R. Schoen and S.-T. Yau, On the Proof of the Positive Mass Conjecture in General Relativity, *Comm. Math. Phys.* **65** (1979), 45–76.
[26] R. Schoen and S.-T. Yau, Positivity of the Total Mass of a General Space-Time, *Phys. Rev. Lett.* **43** (1979), 1457–1459.
[27] R. Schoen and S.-T. Yau, Proof of the Positive Mass Theorem II, *Comm. Math. Phys.* **79** (1981), 231–260.
[28] R. Schoen and S.-T. Yau, The Energy and the Linear Momentum of Space-Times in General Relativity, *Comm. Math. Phys.* **79** (1981), 47–51.
[29] K. P. Tod. *Class. Quant. Grav.* **9** (1992), 1581–1591.
[30] E. Witten, A New Proof of the Positive Energy Theorem, *Comm. Math. Phys.* **80** (1981), 381–402.

ICM 2002 · Vol. II · 273–282

Recent Progress in Kähler Geometry*

Xiuxiong Chen†

Abstract

In recent years, there are many progress made in Kähler geometry. In particular, the topics related to the problems of the existence and uniqueness of extremal Kähler metrics, as well as obstructions to the existence of such metrics in general Kähler manifold. In this talk, we will report some recent developments in this direction. In particular, we will discuss the progress recently obtained in understanding the metric structure of the infinite dimensional space of Kaehler potentials, and their applications to the problems mentioned above. We also will discuss some recent on Kaehler Ricci flow.

2000 Mathematics Subject Classification: 53, 35.
Keywords and Phrases: Extremal Kähler metrics, Kähler-Einstein metrics, Holomorphic vector field, Holomorphic invariant, Kähler Ricci flow.

In the last few years, we have witnessed a rapid progress in Kähler geometry. In particular, the topic related to the existence, to the uniqueness of extremal Kähler metrics, and to obstructions to the existence of such metrics. In this talk, we will give a brief survey of these exciting progress made in this direction.

0.1. Some background

Let (M,ω) be a polarized n-dimensional compact Kähler manifold, where ω is a Kähler form on M. In local coordinates $z_1,\cdots,z_n$, we have

$$\omega = \sqrt{-1}\sum_{i,j=1}^{n} g_{i\overline{j}} d\,z^i \wedge d\,z^{\overline{j}} > 0,$$

where $\{g_{i\overline{j}}\}$ is a positive definite Hermitian matrix function. The Kähler condition requires that ω is a closed positive (1,1)-form. The Kähler metric corresponding to ω is given by

$$g_\omega = \sum_{\alpha,\beta=1}^{n} g_{\alpha\overline{\beta}}\, d\,z^\alpha \otimes d\,z^{\overline{\beta}}.$$

*Paritally supported by NSF research grant DMS-0110321 (2001-2004).
†Department of Mathematics, Princeton University. Department of Mathematics, University of Wisconsin at Madison, USA. E-mail: xiu@math.princeton.edu

For simplicity, in the following, we will often denote by ω the corresponding Kähler metric. The Kähler class of ω is its cohomology class $[\omega]$ in $H^2(M, \mathbb{R})$. It follows from the Hodge-Dolbeault theorem that any other Kähler metric in the same Kähler class is of the form

$$\omega_\varphi = \omega + \sqrt{-1} \sum_{i,j=1}^{n} \frac{\partial^2 \varphi}{\partial z^i \partial z^{\bar{j}}} > 0$$

for some real valued function φ on M.

Given a Kähler metric ω, its volume form is

$$\frac{1}{n!}\omega^n = \left(\sqrt{-1}\right)^n \det\left(g_{i\bar{j}}\right) d\, z^1 \wedge d\, z^{\overline{1}} \wedge \cdots \wedge d\, z^n \wedge d\, z^{\overline{n}}.$$

Its Ricci (curvature) form is:

$$Ric(\omega) = \sqrt{-1} R_{i\bar{j}}\, d\, w_i\, d\, \bar{w}_j = -\sqrt{-1}\partial\bar{\partial}\, \log\, \det\, \omega^n.$$

Note also that $R(\omega) = g^{i\bar{j}} R_{i\bar{j}}$ corresponds to one half times the scalar curvature as it is usually defines in Riemannian geometry. We say that the first Chern class of M is positive of negative definite, if there exists a real valued function ψ on M such that $R_{i\bar{j}} + \frac{\partial^2 \psi}{\partial w_i \partial \bar{w}_j}$ is, respectively, positive of negative definite. A Kähler metric is Kähler-Einstein, if the Ricci form is proportional to the Kähler form by a constant factor. A Kähler metric is called extremal in the sense of E. Calabi [3], if it is a critical point of the functional $\int_M |Ric(\omega)|^2\, \omega^n$, or, equivalently, if the complex gradient vector field of the scalar curvature function $g^{\alpha\bar{\beta}}(\omega) \frac{\partial R(\omega)}{\partial \bar{z}_\beta} \frac{\partial}{\partial z_\alpha}$ is a holomorphic vector field.

0.2. Existence of extremal Kähler metrics

It is well known that a Kähler-Einstein metric satisfies a Monge-Ampere equation

$$\log \det \frac{\omega_\varphi^n}{\omega^n} = -\lambda\, \varphi + h_\omega$$

where $[Ric(\omega)] = \lambda\, [\omega]$ and

$$Ric(\omega) - \lambda\, \omega = i\partial\bar{\partial}\, h_\omega.$$

In Calabi's work in the 1950s, he made conjectures about the existence of Kähler-Einstein metrics on compact Kähler manifolds with definite first Chern class. In 1976, Aubin and Yau independently obtained existence when the first Chern class is negative. Around the same time, Yau proved also the existence of a Kähler-Einstein metric when the first Chern class vanishes. This is a celebrated work; and any Kähler manifold admit such a metric is called "Calabi-Yau" manifold. The positive case remains open, but significant progress has been made in the last two decades. G. Tian proved in [29] the existence of Kähler-Einstein metrics on any complex surface with positive first Chern class and reductive automorphism group.

In 1997, Tian [30] proved that existence of Kähler-Einstein metrics with positive scalar curvature is equivalent to an analytic stability. It remains open how this analytic stability follows from certain algebraic stability in geometric invariant theory.

The construction of complete non-compact Calabi-Yau manifolds has also enjoyed a good deal of success through the work of Calabi, Tian and Yau, Anderson, Kronheimer, LeBrun, Joyce and many others. These non-compact metrics are related to manifolds with G_2 and $Spin(7)$ holonomy, which are important in M-theory.

A lot of effort has also gone into constructing special or explicit examples of Kähler-Einstein metrics and extremal Kähler metrics. The same is true for hyperkaehler metrics as well. Counter examples to the existence of extremal metrics have given by Levine, Burns-De Bartolomeis, and LeBrun.

There has not been much progress made on the existence of extremal metrics in general. One of the possible reasons is the lack of maximum principle for nonlinear equations of 4th order. A general existence result, even in complex surfaces, will be highly interesting.

0.3. Obstructions

In 1983, A. Futaki [19] introduced a complex character $\mathcal{F}(X,[\omega])$ on the complex Lie algebra of all holomorphic vector fields X in M, depending only on the Kähler class $[\omega]$, and show that its vanishing is a necessary condition for the existence of a Kähler-Einstein metric on the manifold. In 1985, E. Calabi[4] generalized Futaki's result to cover the more general case of any extremal Kähler metric: the generalized Futaki invariant of a given Kähler class is zero or not, according to whether any extremal metric in that class has constant scalar curvature or not. S. Bando also obtained some generalizations of the Futaki invariant. More recently, a finite family of obstructions was introduced in [14]. For any holomorphic vector field X inducing the trivial translation on the Albanese torus there exists a complex valued potential function $\theta_{X,\omega}$, uniquely determined up to additive constants, defined by the equation: $L_X\omega = \sqrt{-1}\partial\bar{\partial}\theta_X$(Here L_X denote the Lie derivative along vector field X.). Now, for each $k = 0, 1, \cdots n$, define the functional $\Im_k(X,\omega)$ by[1]

$$\begin{aligned}\Im_k(X,\omega) &= (n-k)\int_M \theta_X\ \omega^n \\ &+ \int_M \left((k+1)\Delta\theta_X\ \mathrm{Ric}(\omega)^k \wedge \omega^{n-k} - (n-k)\ \theta_X\ \mathrm{Ric}(\omega)^{k+1} \wedge \omega^{n-k-1}\right).\end{aligned}$$

Here and elsewhere, Δ_ω denotes the one half times the Laplacian-Beltrami operator of the induced Riemannian structure ω.

The next theorem assures that the above integral gives rise to a holomorphic invariant.

[1]This is a formula for canonical Kähler class. For general Kähler class, see [14].

Theorem 0.1. [14] *The integral $\Im_k(X,\omega)$ is independent of choices of Kähler metrics in the Kähler class $[\omega]$, that is, $\Im_k(X,\omega) = \Im_k(X,\omega')$ so long as the Kähler forms ω and ω' represent the same Kähler class. Hence, the integral $\Im_k(X,\omega)$ is a holomorphic invariant, which will be denoted by $\Im_k(X,[\omega])$. Note that $\Im_0$ is the usual Futaki invariant.*

0.4. Uniqueness of extremal Kähler metrics

We now turn to the uniqueness of extremal metrics. In the 1950s, Calabi used the maximum principle to prove the uniqueness of Kähler-Einstein metrics when the first Chern class is non-positive. In 1987, Mabuchi introduced the "K-energy", which is essentially a potential function for the constant scalar curvature metric equation. Using the K-energy, he and Bando [2] proved that the uniqueness of Kähler-Einstein metric up to holomorphic transformations when the first Chern class is positive. Recently, Tian and X. H. Zhu proved that the uniqueness of Kähler-Ricci Soliton on any Kähler manifolds with positive first Chern class.

Theorem 0.2. [31], [32] *The Kähler Ricci soliton of a Kähler manifold M is unique modulo the automorphism subgroup $Aut_r(M)$; more precisely, if ω_1, ω_2 are two Kähler Ricci solitons with respect to a holomorphic vector field X, i.e., they satisfies*

$$Ric(\omega_i) - \omega_i = \mathcal{L}_X(\omega_i), \qquad \text{where } i = 1, 2. \tag{0.1}$$

Then there are automorphism σ in $Aut^o(M)$ and τ in $Aut_r(M)$ such that $\sigma_*^{-1}X \in \eta_r(M)$ and $\sigma^*\omega_2 = \tau^*\sigma^*\omega_1$, where $\eta_r(M)$ denotes the Lie algebra of $Aut_r(M)$. In fact, $\sigma_*^{-1}X$ lies in the center of $\eta_r(M)$. Moreover, this vector field X is unique up to conjugations.

Following a program of Donaldson (which will be explained in Subsection 0.7), we proved in 1998 [10] that the uniqueness for constant scalar curvature metric in any Kähler class when $C_1 < 0$ along with some other interesting results:

Theorem 0.3. [10] *If the first Chern class is strictly negative, then the extremal Kähler metric is unique in each Kähler class. Moreover, the K energy must have a uniform lower bound if there exists an extremal Kähler metric in that Kähler class.*

Very recently, Donaldson proved a beautiful theorem which states

Theorem 0.4. [18] *For algebraic Kähler class with no non-trivial holomorphic vector field, the constant scalar curvature metric is unique.*

The two theorems overlaps in a lot cases, but mutually non-inclusive.

0.5. Lower bound of the K energy

According to T. Mabuchi and S. Bando[2], the existence of a lower bound of the K energy is a necessary condition for the existence of Kähler-Einstein metrics in the first Chern class. Tian [30] showed that in a Kähler manifold with positive first Chern class and no non-trivial holomorphic fields, the Kähler-Einstein metric

exists if and only if the Mabuchi functional is proper. When the first Chern class is negative, making use of Tian's explicit formulation [30], a simple idea in [9] reduces a lower bound of the K energy to the existence of critical point for the following convex functional:

$$J(\varphi) = -\sum_{p=0}^{n-1} \frac{1}{(p+1)!(n-p-1)!} \int_V \varphi \; Ricci(\omega_0) \wedge \omega_0^{n-p-1} \; (\partial\bar{\partial}\varphi)^p,$$

where $Ricci(\omega_0) < 0$. In complex surfaces, we solves this existence problem completely, which leads to the following interesting result:

Theorem 0.5. [9] *Suppose* $\dim V = 2$ *and* $C_1(V) < 0$. *For any Kähler class* $[\omega_0]$, *if* $2\frac{[-C_1(V)]\cdot[\omega_0]}{[\omega_0]^2}\ [\omega_0] + [C_1(V)] > 0$, *then the K energy has a lower bound in this Kähler class.*

It will be very interesting to generalize this result to higher dimensional Kähler manifold.

0.6. Donaldson's program

Mabuchi defined in [25] a Weil-Petersson type metric on the space of Kähler potentials in a fixed Kähler class. Consider the space of Kähler potentials

$$\mathcal{H} = \{\varphi \mid \omega_\varphi = \omega + \bar{\partial}\partial\varphi > 0, \text{ on } M\}.$$

A tangent vector in $\mathcal{H}$ is just a real valued function in M. For any vector $\psi \in T_\varphi\mathcal{H}$, we define the length of this vector as:

$$\|\psi\|_\varphi^2 = \int_V \psi^2 \; d\,\mu_\varphi.$$

It is easy to see that the geodesic equation for this metric is

$$\varphi''(t) - g_\varphi^{\alpha\bar{\beta}} \frac{\partial\varphi'}{\partial w_\alpha} \frac{\partial\varphi'}{\partial w_{\bar{\beta}}} \;=\; 0,$$

where $g_{\alpha\bar{\beta}_\varphi} = g_{0\,\alpha\bar{\beta}} + \frac{\partial^2\varphi}{\partial w_\alpha \partial w_{\bar{\beta}}} > 0$. It is first observed (cf. Semmes S. [27])that one can complexified the t variable, denoted it by w_{n+1}. Then, the geodesic equation becomes a homogenous complex Monge-Ampere equation:

$$\det \left(g_{0,i\bar{j}} + \frac{\partial^2\varphi}{\partial w_i \overline{\partial w_j}} \right)_{(n+1)(n+1)} \;=\; 0, \qquad \text{on } \Sigma \times M. \tag{0.2}$$

Here $\Sigma = [0,1] \times S^1$. It turns out that we don't need to restrict to this special case. For any Riemann surface Σ with boundary, and for any C^∞ map φ_0 from $\partial\Sigma$ to $\mathcal{H}$, one can always ask the following existence problem:

Question 0.6. (Donaldson[16])For any smooth map $\varphi_0 : \partial\Sigma \to \mathcal{H}$, does there exists a smooth map $\varphi : \Sigma \to \mathcal{H}$ which satisfies the Homogenous Monge-Ampere equation 0.2 such that $\varphi = \varphi_0$ in $\partial\Sigma$?

Theorem 0.7. (X. Chen [10]) *For any smooth map $\varphi_0 : \partial\Sigma \to \mathcal{H}$, there always exists a $C^{1,1}$ map $\varphi : \Sigma \to \overline{\mathcal{H}}$ which solves the Homogenous Monge-Ampere equation 0.2 such that $\varphi = \varphi_0$ in $\partial\Sigma$.*

An important conjecture by Donaldson in [16] was that the space of Kähler potentials is a metric space which is path-connected with respect to this Weil-Petersson metric. This conjecture was complete verified here.

Theorem 0.8. [10] *The space $\mathcal{H}$ is a genuine metric space: the minimal distance between any two Kähler metrics is realized by the unique $C^{1,1}$ geodesic; and the length of this geodesic is positive.*

Collaborating with E. Calabi, we proved the following

Theorem 0.9. [5] *The space $\mathcal{H}$ in a fixed Kähler class is a non-positively curved space in the sense of Alexandrov: Suppose A, B, C are three smooth points in $\mathcal{H}$ and P_λ is a geodesic interpolation point for $0 \leq \lambda \leq 1$: the distance from P to B and C are respectively $\lambda d(B,C)$ and $(1-\lambda)d(B,c)$ [2]. Then the following inequality holds:*

$$d(A,P_\lambda)^2 \leq (1-\lambda)d(A,B)^2 + \lambda d(A,C)^2 - \lambda\cdot(1-\lambda)d(B,C)^2.$$

Theorem 0.10. [5] *Given any two Kähler potentials φ_1 and φ_2 in $\mathcal{H}$ and a smooth curve $\varphi(t), 0 \leq t \leq 1$ which connects them in $\mathcal{H}$. Suppose $\varphi(s,t)$ are the family of curves under the Calabi flow and suppose that $L(s)$ is the length of this curve at time s. Then*

$$\frac{d\,L}{d\,s} = -\int_0^1 \left(\int_M |D\frac{\partial\varphi}{\partial t}|^2_{\varphi(s,t)}\ \omega^n_{\varphi(s,t)} \cdot \sqrt{\int_M |\frac{\partial\varphi}{\partial t}|^2\ \omega^n_{\varphi(s,t)}}^{-\frac{1}{2}} \right) d\,t,$$

where D is the 2nd order Lichernowicz operator. For any smooth function f in V, $D(f) = \sum_{\alpha,\beta=1}^{n} f_{,\alpha\beta} dz^\alpha \otimes dz^\beta$ where $f_{,\alpha\beta}$ is the second covariant derivatives of f.

0.7. The Calabi flow and the Kähler Ricci flow

In a sequence of papers [11], and [14] [12], we develop some new techniques in attacking the convergence problems for the geometric flow, in particular, the Calabi flow and the Kähler Ricci flow. The main ideas are to find a set of new functionals which will be preserved (or decreased) under the flow with a uniform lower bound, then using the principle of concentration to attack the compactness/convergence problem. Following our work [11], M. Struwe [28] gave a more concise proof on Ricci flow and Calabi flow in Riemann surface. This simple idea of using integral estimates in the flow should be able to be applied in other geometric flows.

[2] In affine space, this means $P_\lambda = \lambda\, B + (1-\lambda)\, C$.

0.7.1. The Calabi flow on Riemann surfaces

The Calabi flow is the gradient flow of the K energy and it is a 4th order parabolic equation, proposed by E. Calabi in 1982. Namely, for a given Kähler manifold $(M, [\omega])$, the Calabi flow was defined by

$$\frac{\partial \varphi(t)}{\partial t} = R(\omega_\varphi) - \frac{1}{vol(M)} \int_M R(\omega)\, \omega^n.$$

The local existence for this flow is known, while very little is known for its long term existence since this is a 4th order flow. The only known result is in Riemann surface where Chrusciel proved that the flow converges exponentially fast to a unique constant scalar curvature metric. In 1998 [11], we gave a new proof based on some geometrical integral estimate and concentration compactness principle. Now the challenging question is:

Question 0.11. Does the Calabi flow exists globally for any smooth initial metric?

0.7.2. The Kähler Ricci flow

A Kähler Ricci flowis defined by

$$\frac{\partial}{\partial t}\omega_\varphi = \omega_\varphi - Ric(\omega_\varphi).$$

This flow was first studied by H. D. Cao , following the work of R. Hamilton on the Ricci flow[3]. Cao[6] proved that the flow always exists for all the time along with some other interesting results. It was proved by S. Bando [1] for 3-dimensional Kähler manifolds and by N. Mok [26] for higher dimensional Kähler manifolds that the positivity of bisectional curvature is preserved under the Kähler Ricci flow. The main issue here is the global convergence on manifold with positive bisectional curvature. In the work with Tian, we found a set of new functionals $\{E_k\}_{k=0}^n$ on curvature tensors such that the Ricci flow is the gradient like flow of these functionals. On Kähler-Einstein manifold with positive scalar curvature, if the initial metric has positive bisectional curvature, we can prove that these functionals have a uniform lower bound, via the effective use of Tian's inequality. Consequently, we are able to prove the following theorem:

Theorem 0.12. [14],[12] *Let M be a Kähler-Einstein manifold with positive scalar curvature. If the initial metric has nonnegative bisectional curvature and positive at least at one point, then the Kähler Ricci flow will converge exponentially fast to a Kähler-Einstein metric with constant bisectional curvature.*

The above theorem in complex dimension 1 was proved first by Hamilton [21]. B. Chow [15] later showed that the assumption that the initial metric has positive

[3]The Ricci flow was introduced by R. Hamilton [20] in 1982. There are extensive study in this subject (cf. [22]) since his famous work in 3-dimensonal manifold with positive Ricci curvarure (cf. [22] for further references). Another important geometric flow is the so called "mean curvature flow. " The codimension 1 case was studied extensively by G. Huisken and many others. Recently, there are some interesting progress made in codimension 2 case (cf. [7] [24] for further references).

curvature in S^2 can be removed since the scalar curvature will become positive after finite time anyway.

Corollary 0.13. *The space of Kähler metrics with non-negative bisectional curvature is path-connected.*

Moreover, we can carry over the proof of Theorem 0.12 to a more general case of Kähler orbifolds, for which we will not go into details here. Now the definition of these functionals $E_k = E_k^0 - J_k (k = 0, 1, \cdots, n)$:

Definition 0.14. *For any $k = 0, 1, \cdots, n$, we define a functional E_k^0 on $\mathcal{H}$ by*

$$E_{k,\omega}^0(\varphi) = \frac{1}{vol(M)} \int_M \left(\log \frac{{\omega_\varphi}^n}{\omega^n} - h_\omega \right) \left(\sum_{i=0}^{k} \mathrm{Ric}(\omega_\varphi)^i \wedge \omega^{k-i} \right) \wedge {\omega_\varphi}^{n-k} + c_k,$$

where

$$c_k = \frac{1}{vol(M)} \int_M h_\omega \left(\sum_{i=0}^{k} \mathrm{Ric}(\omega)^i \wedge \omega^{k-i} \right) \wedge \omega^{n-k},$$

and

$$Ric(\omega) - \omega = \sqrt{-1}\partial\bar{\partial} h_\omega, \qquad \text{and} \quad \int_M (e^{h_\omega} - 1)\omega^n = 0.$$

Definition 0.15. *For each $k = 0, 1, 2, \cdots, n-1$, we will define $J_{k,\omega}$ as follows: Let $\varphi(t)$ $(t \in [0, 1])$ be a path from 0 to φ in $\mathcal{H}$, we define*

$$J_{k,\omega}(\varphi) = -\frac{n-k}{vol(M)} \int_0^1 \int_M \frac{\partial \varphi}{\partial t} \left({\omega_\varphi}^{k+1} - \omega^{k+1} \right) \wedge {\omega_\varphi}^{n-k-1} \wedge dt.$$

Put $J_n = 0$ for convenience in notations.

Note that E_0 is the well known K energy function introduced by T. Mabuchi in 1987. Direct computations lead to

Theorem 0.16. *For any $k = 0, 1, \cdots, n$, we have*

$$\begin{aligned} \frac{dE_k}{dt} &= \frac{k+1}{vol(M)} \int_M \Delta_\varphi \left(\frac{\partial \varphi}{\partial t} \right) \mathrm{Ric}(\omega_\varphi)^k \wedge {\omega_\varphi}^{n-k} \\ &\quad - \frac{n-k}{vol(M)} \int_M \frac{\partial \varphi}{\partial t} \left(\mathrm{Ric}(\omega_\varphi)^{k+1} - {\omega_\varphi}^{k+1} \right) \wedge {\omega_\varphi}^{n-k-1}. \end{aligned} \tag{0.3}$$

Here $\{\varphi(t)\}$ is any path in $\mathcal{H}$.

Note that under the Kähler Ricci flow, these functionals essentially decreases! We then prove the derivative of these functionals along a curve of holomorphic automorphisms give rise to a set of holomorphic invariants $\Im_k (k = 0, 1, \cdots, n)$ (cf. Theorem 0.1). In case of Kähler-Einstein manifolds, all these invariants vanishes. This give us freedom to re-adjust the flow so that the evolving Kähler potentials are perpendicular to the first eigenspace of a fixed Kähler-Einstein metric. Then we will be able to show that the evolved volume form has a uniform lower bound. From this point on, the boot-strapping process will give us necessary estimates to obtain global convergence.

0.8. Some new result with G. Tian

In 2001, Donaldson proved the following

Theorem 0.17. [17] *(Openness) For any smooth solution to the geodesic equation with a disc domain, there are always exists a smooth solution to the geodesic equation if we perturb the boundary data in a small open set (of the given boundary data).*

This is somewhat surprising result since it is very hard to deform any solution of a homogenous Monge-Ampere equation even locally. However, Donaldson was able to make clever use of the Fredholm theory of holomorphic discs with totally real boundary in his proof. Then the problem of closed-ness becomes very important in light of this theorem. Tian and I are able to establish the closed-ness in this case.

Theorem 0.18. [13] *(Closure property) The defomation of geodesic solution in the preceding theorem is indeed closed, provided we allow solution to be smooth almost everywhere.*

This is a deep theorem and we will not go into detail here due to the expository nature of this talk. However, this theorem, along with the ideas of proof, shall have implication in both geometry and other Monge-Ampere type equation in the future.

References

[1] S. Bando. On the three dimensional compact Kähler manifolds of nonnegative bisectional curvature. *J. D. G.*, 19:283–297, 1984.

[2] S. Bando and T. Mabuchi. *Uniquness of Einstein Kähler metrics modulo connected group actions.* In *Algebraic Geometry*, Advanced Studies in Pure Math., 1987.

[3] E. Calabi. Extremal Kähler metrics. In *Seminar on Differential Geometry*, volume 16 of *102*, 259–290. Ann. of Math. Studies, University Press, 1982.

[4] E. Calabi. Extremal Kähler metrics, II. In *Differential geometry and Complex analysis*, 96–114. Springer, 1985.

[5] E. Calabi and X. X. Chen. Space of Kähler metrics (II), 1999. to appear in J.D.G.

[6] H. D. Cao. Deformation of Kähler metrics to Kähler-Einstein metrics on compact Kähler manifolds. *Invent. Math.*, 81:359–372, 1985.

[7] J.Y. Chen and J.Y. Li. *mean curvature flow of surfaces in 4-manifolds*, 2000. Adv. in Math. (to appear).

[8] J.Y. Chen and J.Y. Li. *quaternionic maps between hyperkähler manifolds*, 2000. J. D. G.

[9] X. X. Chen. On lower bound of the Mabuchi energy and its application. *International Mathematics Research Notices*, 12, 2000.

[10] X. X. Chen. Space of Kähler metrics. *Journal of Differential Geometry,*, 56:189–234, 2000.

[11] X. X. Chen. Calabi flow in Riemann surface revisited: a new point of views. (6):276–297, 2001. "International Mathematics Research Notices".

[12] X. X. Chen and G. Tian. Ricci flow on Kähler-Einstein manifolds, 2000. Submitted to Annals of Mathematics.
[13] X. X. Chen and G. Tian. Space of Kähler metrics (III), 2000. preprint.
[14] X. X. Chen and G. Tian. Ricci flow on complex surfaces, 2002. *Inventiones mathemticae.*
[15] B. Chow. The Ricci flow on the 2-sphere. *J. Diff. Geom.*, 33:325–334, 1991.
[16] S.K. Donaldson. Symmetric spaces, kahler geometry and Hamiltonian dynamics. *Amer. Math. Soc. Transl. Ser. 2, 196*, 13–33, 1999. Northern California Symplectic Geometry Seminar.
[17] S.K. Donaldson. Holomorphic Discs and the complex Monge-Ampere equation, 2001. to appear in Journal of Sympletic Geometry.
[18] S.K. Donaldson. Scalar curvature and projective embeddings, I, 2001. to appear in Journal of Differential Geometry.
[19] A. Futaki. An obstruction to the existence of Einstein Kähler metrics. *Inv. Math. Fasc.*, 73(3):437–443, 1983.
[20] R. Hamilton. Three-manifolds with positive Ricci curvature. *J. Diff. Geom.*, 17:255–306, 1982.
[21] R. Hamilton. The Ricci flow on surfaces. *Contemporary Mathematics*, 71:237–261, 1988.
[22] R. Hamilton. *The formation of singularities in the Ricci flow*, volume II. Internat. Press, 1993.
[23] J.Y. Li J.Y. Chen and G. Tian. *two dimensional graphs moving by mean curvature flow*, 2000. preprint.
[24] Wang M. T. Mean curvature flow of surfaces in einstein four-manifolds. *J. D. Geometry*, 57(2):301–338, 2001.
[25] T. Mabuchi. Some Sympletic geometry on compact kähler manifolds I. *Osaka, J. Math.*, 24:227–252, 1987.
[26] N. Mok. The uniformization theorem for compact Kähler manifolds of nonnegative holomorphic bisectional curvature. *J. Differential Geom.*, 27:179–214, 1988.
[27] S. Semmes. Complex monge-ampere and sympletic manifolds. *Amer. J. Math.*, 114:495–550, 1992.
[28] M. S. Struwe. Curvature flows on surfaces, August 2000. priprint.
[29] G. Tian. On Calabi's conjecture for complex surfaces with positive first chern class. *Invent. Math.*, 101(1):101–172, 1990.
[30] G. Tian. Kähler-Einstein metrics with positive scalar curvature. *Invent. Math.*, 130:1–39, 1997.
[31] G. Tian and X. H. Zhu. Uniqueness of kähler-Ricci Soliton, 1998. priprint.
[32] G. Tian and X. H. Zhu. A new holomorphic invariant and uniqueness of Kähler-Ricci Soliton, 2000. priprint.

ICM 2002 · Vol. II · 283–291

On the Schrödinger Flows

Weiyue Ding*

Abstract

We present some recent results on the existence of solutions of the Schrödinger flows, and pose some problems for further research.

2000 Mathematics Subject Classification: 53C44, 35Q55.
Keywords and Phrases: Schrödinger equation, Hamiltonnian flow, Kähler manifold.

1. Introduction

Recently the research on so-called Schrödinger flow (or Schrödinger map [1]-[4]) has been carried out by several authors. This is an infinite-dimensional Hamiltonian flow defined on the space of mappings from a Riemannian manifold (M, g) into a Kähler manifold (N, J, h), where g is the Riemannian metric on M, and h is the Kähler metric on N, with J being the complex structure on N. This flow is defined by the following equation

$$u_t = J(u)\tau(u), \tag{1.1}$$

where $\tau(u)$ is the so-called tension field well-known in the theory of harmonic maps. In local coordinates, $\tau(u)$ is given by

$$\tau(u)^i = \Delta_M u^i - g^{\alpha\beta}\Gamma^i_{jk}(u)\frac{\partial u^j}{\partial x^\alpha}\frac{\partial u^k}{\partial x^\beta}.$$

Here Δ_M is the Laplace-Beltrami operator on M and Γ^i_{jk} are the Christoffel symbols of the Riemannian connection on N. Obviously, the Schrödinger flows preserves the energy $E(u)$ of mapping u, i.e. $E(u(t)) \equiv E(u(0))$, where

$$E(u) = \frac{1}{2}\int_M g^{\alpha\beta}h_{jk}(u)\frac{\partial u^j}{\partial x^\alpha}\frac{\partial u^k}{\partial x^\beta}dM.$$

*Peking Univ. and AMSS, CAS, China. E-mail: dingwy@math.pku.edu.cn

Schrödinger flows are related to various theories in mechanics and physics. A well-known and important example is the so-called Heisenberg spin chain system (also called ferromagnetic spin chain system [7]). This is just the Schrödinger flow into S^2. Consider S^2 as the unit sphere in R^3, then the equation for the system is given by

$$u_t = u \times \Delta u.$$

Note that, for a mapping u from M into S^2,

$$J(u) = u\times : T_uS^2 \longrightarrow T_uS^2$$

is the standard complex structure on S^2, and the tension field of the map u into S^2 is given by $\tau(u) = \Delta u + |\nabla u|^2 u$. So, we have $u \times \Delta u = J(u)\tau(u)$. Another interesting example of the Schrödinger flow is the anisotropic Heisenberg spin chain system, i.e. the Schrödinger flow into Poincaré disk $H(-1)$.

Comparing to other geometric nonlinear evolutionary systems, such as the heat flow of harmonic maps (parabolic system) and wave maps (hyperbolic system), the study of Schrödinger flows is still at the beginning stage. There are some remarkable results on the existence of solutions for certain specific cases. E.g. for the Heisenberg spin chain system ($N = S^2$), Zhou et. al. [9] proved the global existence for $M = S^1$, and Sulem et. al. [10] proved the local existence for $M = R^m$. There are some more recent works, see [1], [3] and [11]. For the general case, however, it turns out that even local existence is hard to prove. In this respect, a recent result obtained by Youde Wang and this author ([4]) states

Theorem *Let (M, g) be a closed Riemannian manifold of dimension m, and let (N, J, h) is a closed Kähler manifold. If m_0 is the smallest integer greater than $m/2$ (i.e. $m_0 = [m/2] + 1$), and $u_0 \in W^{k,2}(M, N)$ for any $k \geq m_0 + 3$, then the initial value problem for (1.1) with initial value u_0 has a unique local solution. Moreover, if $u_0 \in C^\infty(M, N)$, the local solution is C^∞ smooth.*

We remark that, the maximal existence time of the local solution in the above result, depends only on the W^{m_0+1}-norm of the initial map u_0 for any k. This is why we can get local existence in the C^∞ case. Also, for the existence part, the regularity of u_0 can be lowered to $W^{k,2}$ with $k \geq m_0 + 1$, however we do not know how to get the uniqueness if $k < m_0 + 3$.

In the following, we give a description of the proof of the above Theorem in Section 2 and 3. Then, in Section 4, we pose some important problems for future research of the Schrödinger flows.

2. Some inequalities for Sobolev section norms of maps

Let $\pi : E \longrightarrow M$ be a Riemannian vector bundle over M. Then we have the bundle $\Lambda^p T^*M \otimes E \longrightarrow M$ over M which is the tenser product of the bundle E and the induced p-form bundle over M, where $p = 1, 2, \cdots, \dim(M)$. We define $\Gamma(\Lambda^p T^*M \otimes E)$ as the set of all smooth sections of $\Lambda^p T^*M \otimes E \longrightarrow M$. There

exists a induced metric on $\Lambda^p T^*M \otimes E \longrightarrow M$ from the metric on T^*M and E such that for any $s_1, s_2 \in \Gamma(\Lambda^p T^*M \otimes E)$

$$\langle s_1, s_2 \rangle = \sum_{i_1<i_2<\cdots<i_p} \langle s_1(e_{i_1}, \cdots, e_{i_p}), s_2(e_{i_1}, \cdots, e_{i_p}) \rangle,$$

where $\{e_i\}$ is an orthonormal local frame of TM. We define the inner product on $\Gamma(\Lambda^p T^*M \otimes E)$ as follows

$$(s_1, s_2) = \int_M \langle s_1, s_2 \rangle(x) dM = \int_M \langle s_1, s_2 \rangle(x) * 1.$$

The Sobolev space $L^2(M, \Lambda^p T^*M \otimes E)$ is the completion of $\Gamma(\Lambda^p T^*M \otimes E)$ with respect to the above inner product $(\cdot, \cdot)$, we may also define analogously the Sobolev spaces $H^{k,r}(M, \Lambda^p T^*M \otimes E)$ or $H^{k,r}(M, E)$. Let ∇ be the covariant differential induced by the metric on E, then we can take the completion of the smooth sections of E in the norm,

$$||s||_{k,r} = ||s||_{H^{k,r}(M,E)} = (\sum_{i=0}^{k} \int_M |\nabla^i s|^r dM)^{\frac{1}{r}}.$$

We call the above Sobolev spaces as the bundle-valued Sobolev spaces.

In [4] We establish the following interpolation inequality for sections on vector bundles, which was proved for functions on $\mathbb{R}^m$ by Gagliardo and Nirenberg, and for functions on Riemannian manifolds by Aubin ([8]).

Lemma 2.1 *Let M be a compact Riemannian manifold with $\dim(M) = m$ and E be a Riemannian vector bundle over M. Let q, r be real numbers $1 \leq q, r \leq \infty$ and j, m integers $0 \leq j \leq n$. Then there exists a constant $C(M)$ depending m, n, j, q, r and a, and on M, but not depending on the choice of metrics on E, such that for all $s \in C^\infty(E)$:*

$$\|\nabla^j s\|_{L^p} \leq C(M) \|s\|^a_{H^{n,r}} \|s\|^{1-a}_{L^q}, \tag{2.1}$$

where

$$\frac{1}{p} = \frac{j}{m} + a\left(\frac{1}{r} - \frac{n}{m}\right) + (1-a)\frac{1}{q},$$

for all a in the interval $\frac{j}{n} \leq a \leq 1$, for which p is non-negative. If $r = \frac{m}{n-j} \neq 1$, then the above interpolation inequality is not true for $a = 1$.

The so-called Sobolev section norms of mapping $u \in C^\infty(M, N)$, where M is a closed Riemannian manifold, is defined as the Sobolev section norms of ∇u where ∇u is regarded as a section on the bundle $u^*(TN) \otimes T^*M$. Then with $s = \nabla u$, we have by Lemma 2.1,

$$\|\nabla^{j+1} u\|_{L^p} \leq C \|\nabla u\|^a_{H^{k,q}} \|\nabla u\|^{1-a}_{L^r}, \tag{2.2}$$

where the constants in (2.1) satisfy the conditions of Lemma 2.1. Obviously, the $H^{k,2}$ norm of maps $u \in C^\infty(M, N)$ is nonlinear with respect to u.

In order to prove Theorem we need to consider the problem of comparing the $W^{k,2}$ norm with $H^{k,2}$ norm of maps $u \in C^\infty(M,N)$ (i.e. Sobolev section norm). We assume that M is a closed Riemannian manifold and N is a compact Riemannian manifold with or without boundary. It will be convenient to imbed N isometrically into some Euclidean space $\mathbb{R}^K$, and consider N as a compact submanifold of $\mathbb{R}^K$. Then the map u can be represented as $u = (u^1, \cdots, u^K)$ with u^i being globally defined functions on M. The we have

$$\|u\|^2_{W^{k,2}} = \sum_{i=0}^{k} \|D^i u\|^2_{L^2},$$

where

$$\|D^i u\|^2_{L^2} = \sum_{|\mathbf{a}|=i} \|D_{\mathbf{a}} u\|^2_{L^2},$$

and D denotes the covariant derivative for functions on M. The $H^{k,2}$ norm of u is defined similarly, only we need to replace D by ∇, where ∇ is the covariant derivative for sections of the bundle $u^*(TN)$ over M(For simplicity we also write $\nabla u = Du$). In [4] Ding and Wang obtained the following lemma.

Lemma 2.2 *Assume that $k > m/2$. Then there exists a constant C=C(N,k) such that for all $u \in C^\infty(M,N)$,*

$$\|Du\|_{W^{k-1,2}} \leq C \sum_{t=1}^{k} \|\nabla u\|^t_{H^{k-1,2}}, \tag{2.3}$$

and

$$\|\nabla u\|_{H^{k-1,2}} \leq C \sum_{t=1}^{k} \|Du\|^t_{W^{k-1,2}}. \tag{2.4}$$

3. The proof of theorem

In this section we prove the local existence of smooth solutions for the initial value problem of the Schrödinger flow

$$\begin{cases} u_t = J(u)\tau(u), \\ u(\cdot,0) = u_0 \in C^\infty(M,N). \end{cases} \tag{3.1}$$

We need to employ an approximate procedure and solve first the following perturbed problem

$$\begin{cases} u_t = \epsilon\tau(u) + J(u)\tau(u), \\ u(\cdot,0) = u_0 \in C^\infty(M,N), \end{cases} \tag{3.2}$$

where $\epsilon > 0$ is a small number.

The advantage of (3.2) is that the equation with $\epsilon > 0$ is uniformly parabolic. Hence the initial value problem has a unique smooth solution $u_\epsilon \in C^\infty(M \times [0,T_\epsilon], N)$ for some $T_\epsilon > 0$. The problem is then to obtain a uniform positive

lower bound T of T_ϵ, and uniform bounds for various norms of $u_\epsilon(t)$ in suitable spaces for t in the time interval $[0,T)$. (Since we shall use L^2 estimates, the norms are $W^{k,2}(M,N)-$norms for all positive integer k.) Once we get these bounds it is clear that the u_ϵ subconverge to a smooth solution of (3.1) as $\epsilon \to 0$.

Now let $u = u_\epsilon$ be a solution of (3.2), then it is easy to see that the energy $E(u(t))$ is uniformly bounded for $t \in [0, T_\epsilon)$, i.e.

$$E(u(t)) \leq E(u_0). \tag{3.3}$$

In the following we will make estimations on L^2-norms of all covariant derivatives $\nabla^k u$ $(k = 2, 3, \cdots)$.

Lemma 3.1 *Let $m_0 = [m/2]+1$, where $[q]$ denotes the integral part of a positive number q, and let $u_0 \in C^\infty(M,N)$. There exists a constant $T = T(\|u_0\|_{H^{m_0+1,2}}) > 0$, independent of $\epsilon \in [0,1]$, such that if $u \in C^\infty(M \times [0,T_\epsilon])$ is a solution of (3.1) with $\epsilon \in (0,1]$ then*

$$T_\epsilon \geq T(\|\nabla u_0\|_{H^{m_0,2}})$$

and

$$\|\nabla u(t)\|_{H^{k,2}} \leq C(k, \|\nabla u_0\|_{H^{k,2}}) \quad t \in [0,T]$$

for all $k \geq m_0$.

Proof Fix a $k \geq m_0$, and let l be any integer with $1 \leq l \leq k$. Suppose that $\mathbf{a}$ be a multi-index of length l, i.e. $\mathbf{a} = (a_1, \cdots, a_l)$. Then we have for $t \leq T_\epsilon$

$$\frac{1}{2}\frac{d}{dt}\|\nabla_{\mathbf{a}}\nabla_i u\|_{L^2}^2 = \int_M \langle \nabla_{\mathbf{a}}\nabla_i u, \nabla_t \nabla_{\mathbf{a}}\nabla_i u\rangle. \tag{3.4}$$

Exchanging the order of covariant differentiation we have (cf. [9])

$$\nabla_t\nabla_{\mathbf{a}}\nabla_i u = \nabla_{\mathbf{a}}\nabla_i\nabla_t u + \sum \nabla_{\mathbf{b}} R(u)(\nabla_{\mathbf{c}} u, \nabla_{\mathbf{d}}\nabla_t u)\nabla_{\mathbf{e}}\nabla_i u,$$

where the sum is over all multi-indexes $\mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}$ with possible zero lengths, except that $|\mathbf{c}| > 0$ always holds, such that

$$(\mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}) = \sigma(\mathbf{a})$$

is a permutation of $\mathbf{a}$. Noting that we may replace $\nabla_t u$ in the terms of the summation by the right hand side of equation (3.2), the above identity can be rewritten as

$$\nabla_t\nabla_{\mathbf{a}}\nabla_i u = \nabla_{\mathbf{a}}\nabla_i\nabla_t u + Q \tag{3.5}$$

with

$$|Q| \leq C(l,M)\sum |\nabla^{j_1} u| \cdots |\nabla^{j_s} u| \tag{3.6}$$

where the summation is over all $(j_1, \cdots, j_s)$ satisfying

$$j_1 \geq j_2 \geq \cdots \geq j_s, \;\; l+1 \geq j_i \geq 1, \;\; j_1 + \cdots + j_s = l+3, \;\; s \geq 3. \tag{3.7}$$

For the first term in the right hand side of (3.5), we may use the equation (3.2) to get

$$\begin{aligned}\nabla_{\mathbf{a}}\nabla_i\nabla_t u &= \nabla_{\mathbf{a}}\nabla_i(\epsilon\tau(u)+J(u)\tau(u))\\ &= \epsilon\nabla_{\mathbf{a}}\nabla_i\nabla_k\nabla_k u + J(u)\nabla_{\mathbf{a}}\nabla_i\nabla_k\nabla_k u \qquad (3.8)\end{aligned}$$

where we have used the integrability of the complex structure J of the Kähler manifold N. By exchanging the orders of covariant differentiation as above, we get from (3.5) and (3.8)

$$\nabla_t\nabla_{\mathbf{a}}\nabla_i u = \epsilon\nabla_k\nabla_k\nabla_{\mathbf{a}}\nabla_i u + J(u)\nabla_k\nabla_k\nabla_{\mathbf{a}}\nabla_i u + Q$$

where Q satisfies (3.6-3.7). Substituting this into (3.4) and integrating by part we then have

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|\nabla_{\mathbf{a}}\nabla_i u\|^2_{L^2}\\ =\ &\int_M(-\epsilon|\nabla\nabla_{\mathbf{a}}\nabla_i u|^2 - \langle\nabla_k\nabla_{\mathbf{a}}\nabla_i u, J(u)\nabla_k\nabla_{\mathbf{a}}\nabla_i u\rangle + \langle\nabla_{\mathbf{a}}\nabla_i u, Q\rangle).\end{aligned}$$

Note that the first integrand is non-positive and the second vanishes, so we have by (3.6)

$$\frac{d}{dt}\|\nabla_{\mathbf{a}}\nabla_i u\|^2_{L^2} \le C(l,M)\sum\int_M|\nabla^{l+1}u||\nabla^{j_1}u|\cdots|\nabla^{j_s}u|,$$

and consequently

$$\frac{d}{dt}\|\nabla^{l+1}u\|^2_{L^2} \le C(l,M)\sum\int_M|\nabla^{l+1}u||\nabla^{j_1}u|\cdots|\nabla^{j_s}u|, \qquad (3.9)$$

where the summation is over all $(j_1,\cdots j_s)$ satisfying (3.7).

To treat the integrals in the summation of (3.9), i.e.

$$I = \int_M|\nabla^{l+1}u||\nabla^{j_1}u|\cdots|\nabla^{j_s}u|, \qquad (3.10)$$

we need the following lemmas which can be proved by applying Lemma 2.1, the Hölder inequality and some combination techniques. Especially, the proof of Lemma 3.3 is slightly tricky, for details we refer to [4].

Lemma 3.2 *Let I be the integral (3.10), where $(j_1,\cdots,j_s)$ satisfy (3.7). If $1\le l\le m_0$, then there exists a constant $C=C(M,l)$ such that*

$$I \le C\|\nabla u\|^A_{H^{m_0,2}}\|\nabla u\|^B_{L^2}\|\nabla^{l+1}u\|_{L^2},$$

where $A=[l+3+(m/2-1)s-m/2]/m_0$ and $B=s-A$.

Lemma 3.3 *Assume $l>m_0$. Then there exists a constant $C=C(M,l)$ such that*
(1) if $j_1=l+1$,

$$I \le C\|\nabla^{l+1}u\|^2_{L^2}\|\nabla u\|^{m/m_0}_{H^{m_0,2}}\|\nabla u\|^{2-m/m_0}_{L^2}.$$

(2) if $j_1 \le l$,

$$I \le C(1 + \|\nabla u\|^2_{H^{l,2}})(1 + \|\nabla u\|^A_{H^{l-1,2}})$$

where $A = A(m,l)$.

Now, return to the proof of Lemma 3.1. We first consider the case $1 \le l \le m_0$ in (3.9). Then Lemma 3.2 together with (3.3) leads to

$$\frac{d}{dt}\|\nabla u\|_{H^{m_0,2}} \le C \sum_{l=1}^{m_0} \sum_{s=3}^{l+3} \|\nabla u\|^{A(s,l)}_{H^{m_0,2}},$$

where

$$A(s,l) = [l + 3 + (m/2 - 1)s - m/2]/m_0.$$

If we let $f(t) = \|\nabla u(t)\|_{H^{m_0,2}} + 1$, then we have

$$f' \le C f^{A_0}, \quad f(0) = \|\nabla u_0\|_{H^{m_0,2}} + 1, \tag{3.11}$$

where $A_0 = \max\{A(s,l) : 3 \le s \le l+3, 1 \le l \le m_0\}$. The constant C in (3.11) depends only on m_0, M and N. It follows from (3.11) that there exists $T = T(N, \|\nabla u_0\|_{W^{m_0,2}}) > 0$ and $K_0 > 0$ such that

$$\|\nabla u(t)\|_{H^{m_0,2}} \le K_0, \quad t \in [0,T]. \tag{3.12}$$

For any $k > m_0$, we need to consider the case $m_0 < l \le k$ in (3.9). Lemma 3.3, (3.3) and (3.12) then imply

$$\frac{d}{dt}\|\nabla u\|^2_{H^{k,2}} \le C(1 + \|\nabla u\|^2_{H^{k,2}})(1 + \|\nabla u\|^A_{H^{k-1,2}}). \tag{3.13}$$

For $k = m_0 + 1$, we see from (3.12) that the summation in (3.13) is bounded since $k - 1 = m_0$. Then, since (3.13) is a linear differential inequality for $\|\nabla u\|^2_{H^{k,2}}$, there exists a constant $K_1 > 0$ such that

$$\|\nabla u(t)\|_{H^{m_0+1,2}} \le K_1, \quad t \in [0,T]. \tag{3.14}$$

It now is clear that one can show inductively using (3.13) the existence of $K_i > 0$ for any $i \ge 1$ such that

$$\|\nabla u(t)\|_{H^{m_0+i,2}} \le K_i, \quad t \in [0,T]. \tag{3.15}$$

Since we assume M is compact, consequently $\|u(t)\|_{L^\infty}$ is uniformly bounded for $t \in [0,T]$.

It is easy to find that the solution to (3.2) with $\epsilon \in (0,1)$ must exist on the time interval $[0,T]$. Otherwise, we always extend the time interval of existence to cover $[0,T]$, i.e., we always have $T_\epsilon \ge T$. Thus, Lemma 3.1 has been proved.

Proof of Theorem First, we would like to mention that N is always regarded as an embedded submanifold of $\mathbb{R}^K$. If $u_0 : M \longrightarrow N$ is C^∞, then, Lemma 3.1 claims that the initial value problem (3.2) admits a unique smooth solution u_ϵ

which satisfies the estimates in Lemma 3.1. It follows from Proposition 2.2 that, for any $k > 0$ and $\epsilon \in (0, 1]$, there holds

$$\max_{t\in[0,T]} \|u_\epsilon\|_{W^{k,2}(M)} \leq C_k(M, u_0),$$

where $C_k(M, u_0)$ does not depend on ϵ. Hence, by sending $\epsilon \longrightarrow 0$ and applying the embedding theorem of Sobolev spaces to u, we have $u_\varepsilon \longrightarrow u \in C^k(M \times [0, T], N)$ for any k. It is very easy to check that u is a solution to the initial value problem (3.1). The uniqueness was addressed in Proposition 2.1 in [1].

Finally, if $u_0 : M \longrightarrow N$ is not C^∞, but $u_0 \in W^{k,2}(M, N)$, we may always select a sequence of C^∞ maps from M into N, denoted by u_{i0}, such that

$$u_{i0} \longrightarrow u_0 \ \text{ in } \ W^{k,2}, \quad \text{ as } \ i \to \infty.$$

This together with the definition of covariant differential leads to

$$\|\nabla u_{i0}\|_{H^{k-1,2}} \longrightarrow \|\nabla u_0\|_{H^{k-1,2}}, \quad \text{ as } \ i \to \infty.$$

Thus, there exists a unique, smooth solution u_i, defined on time interval $[0, T_i]$, of the Cauchy problem (3.1) with u_0 replaced by u_{i0}. Furthermore, it is not difficult to see from the arguments in Lemma 3.1 that if i is large enough, then there exists a uniform positive lower bound of T_i, denoted by T, such that the following holds uniformly with respect to large enough i:

$$\sup_{t\in[0,T]} \|\nabla u_i(t)\|_{H^{k-1,2}} \leq C(T, \|\nabla u_0\|_{H^{k-1,2}}).$$

It follows from Lemma 2.2 and the last inequality that

$$\sup_{t\in[0,T]} \|Du_i(t)\|_{W^{k-1,2}} \leq C'(T, \|Du_0\|_{W^{k-1,2}}),$$

where D denotes the covariant derivative for functions on M. Therefore, there exists a $u \in L^\infty([0, T], W^{k-1,2}(M, N))$ such that

$$u_i \longrightarrow u \ \text{ [weakly*] in } \ L^\infty([0, T], W^{k,2}(M, N))$$

upon extracting a subsequence and re-indexing if necessary. It is easy to verify that u is a strong solution to (3.1) (see [4]).

Remark For the Schrödinger flow from an Euclidean space into a Kähler manifold, in [4] we obtained similar local existence results.

4. Some problems

1. For the one-dimensional case, i.e. $\dim M = 1$, we conjecture the Schrödinger flows should exist globally whenever the target N is a compact Kähler manifold. This is still open, and is supported by the result with N being Hermitian locally symmetric ([11]).

The result by Terng and Uhlenbeck [2] shows that for some special targets (e.g. complex Grassmannians), the Schrödinger flows are bi-Hamiltonnian integrable systems. In their work, they assume that $M = R^1$, and their result can be generalized to compact Hermitian symmetric spaces (cf. [12]). An interesting open problem is, for these special targets, whether or not the Schrödinger flows are bi-Hamiltonnian systems if $M = S^1$.

2. For higher dimensional cases, i.e. $\dim M \geq 2$, we believe that the Schrödinger flow may develop finite-time singularities. There are however no such examples known by now.

3. All present results in the study of the Schrödinger flows depend on the global estimates for the solutions. We do not know if one can find some kind of *local estimates* for the solutions. It has been well known from the research of various geometric flows that local estimates are important for the analysis of singularities. It is therefore desirable to develop some new methods to attack the question before any serious advance can be made for the study of the Schrödinger flows.

References

[1] W. Y. Ding and Y. D. Wang, Schrödinger flows of maps into symplectic manifolds, *Science in China A*, **41(7)**(1998), 746–755.

[2] C. T. Terng and K. Uhlenbeck, Schrödinger flows on Grassmannians, math. DG/9901086.

[3] N. Chang, J. Shatah and K. Uhlenbeck, Schrödinger maps, *Comm. Pure Appl. Math.*, **53**(2000), 590–602.

[4] W. Y. Ding and Y. D. Wang, Schrödinger flows into Kähler manifolds, *Science in China A*, **44(11)**(2001), 1446–1464.

[5] L. D. Landau and E. M. Lifshitz, On the theory of the dispersion of magnetic permeability in ferromagnetic bodies, *Phys. Z. Sowj.* **8**(1935), 153; reproduced in *Collected Papers of L. D. Landau*, Pergaman Press, New York, 1965, 101–114.

[6] T. Aubin, *Nonlinear Analysis on Manifolds. Monge-Ampère Equations*, Springer-Verlag, Berlin-Heidelberg-New York, 1982.

[7] J. Eells and L. Lemaire, Another report on harmonic maps, *Bull. London Math. Soc.*, **20** (1988), 385–524.

[8] L. Fadeev and L. A. Takhatajan, *Hamiltonian Methods in the Theory of Solitons*, Springer-Verlag, Berlin-Heidelberg-New York, 1987.

[9] Y. Zhou, B. Guo and S. Tan, Existence and uniqueness of smooth solution for system of ferromagnetic chain, *Science in China A*, **34**(1991), 257–266.

[10] P. Sulem, C. Sulem and C. Bardos, On the continuous limit for a system of classical spins, *Commun. Math. Phys.*, **107** (1986), 431-454.

[11] P. Pang, H. Wang, Y. D. Wang, Schrödinger flow on Hermitian locally symmetric spaces, to appear in *Comm. Anal. Geom.* .

[12] B. Dai, Ph. D. dissertation of NUS (2002).

ICM 2002 · Vol. II · 293–302

Differential Geometry via Harmonic Functions

P. Li*

Abstract

In this talk, I will discuss the use of harmonic functions to study the geometry and topology of complete manifolds. In my previous joint work with Luen-fai Tam, we discovered that the number of infinities of a complete manifold can be estimated by the dimension of a certain space of harmonic functions. Applying this to a complete manifold whose Ricci curvature is almost non-negative, we showed that the manifold must have finitely many ends. In my recent joint works with Jiaping Wang, we successfully applied this general method to two other classes of complete manifolds. The first class are manifolds with the lower bound of the spectrum $\lambda_1(M) > 0$ and whose Ricci curvature is bounded by

$$Ric_M \geq -\frac{m-2}{m-1}\lambda_1(M).$$

The second class are stable minimal hypersurfaces in a complete manifold with non-negative sectional curvature. In both cases we proved some splitting type theorems and also some finiteness theorems.

2000 Mathematics Subject Classification: 53C21, 58J05.
Keywords and Phrases: Harmonic function, Ricci curvature, Minimal hypersurface, Parabolic manifold.

1. Introduction

In 1992, the author and Luen-fai Tam [12] discovered a general method to determine if a complete, non-compact, Riemannian manifold have finitely many ends. An end is simply defined to be an unbounded component of the compliment of a compact set in the manifold. If the number of ends is finite, their technique also provides an estimate on the number of ends. In particular, they applied this method to prove that a certain class of manifolds must have finitely many ends.

*Department of Mathematics, University of California, Irvine, CA 92697-3875, USA. E-mail: pli@math.uci.edu

Theorem 1 (Li-Tam). *Let M^m be a complete, non-compact, manifold with*

$$Ric_M(x) \geq -k(r(x)),$$

where $k(r)$ is a continuous non-increasing function satisfying

$$\int_0^\infty r^{m-1}\, k(r)\, dr < \infty.$$

Then there exists a constant $0 < C(m,k) < \infty$ depending only on m and k, such that, M has at most $C(m,k)$ number of ends.

Since a manifold with non-negative Ricci curvature will satisfy the hypothesis, this theorem can be viewed as a perturbed version of the splitting theorem [4] of Cheeger-Gromoll. A weaker version of the above theorem for manifolds with non-negative Ricci curvature outside a compact set was also independently proved by Cai [1].

In some recent work of Jiaping Wang and the author, they successfully applied the general theory of determining the number of ends to other situations. The purpose of this note is to give a quick overview of the theory and its applications to manifolds with positive spectrum and minimal hypersurfaces.

2. General theory

Throughout this article, we will assume that $(M^m,\ ds^2_M)$ is an m-dimensional, complete, non-compact Riemannian manifold without boundary. In terms of local coordinates $(x_1, x_2, \ldots, x_m)$, if the metric is given by

$$ds^2_M = g_{ij}\, dx_i\, dx_j,$$

then the Laplacian is defined by

$$\Delta = \frac{1}{\sqrt{g}} \frac{\partial}{\partial x_i} \left(g_{ij} \sqrt{g} \frac{\partial}{\partial x_j} \right),$$

where $(g^{ij}) = (g_{ij})^{-1}$ and $g = \det(g_{ij})$. A function is said to be *harmonic* on M if it satisfies the Laplace equation

$$\Delta f(x) = 0$$

for all $x \in M$.

In order to state the general theorem, it is necessary for us to define the following spaces.

Definition 1. *Let*

$$\mathcal{H}_D(M) = \{f \,|\, \Delta f = 0,\ ||f||_\infty < \infty,\ \int_M |\nabla f|^2 < \infty\}$$

be the space of bounded harmonic functions with finite Dirichlet integral defined on M.

Definition 2. *Let*

$$\mathcal{H}_+(M) = \langle \{f \mid \Delta f = 0,\ f > 0\} \rangle$$

be the space spanned by the set of positive harmonic functions defined on M.

Definition 3. *Let*

$$\mathcal{H}'(M) = \langle \{f \mid \Delta f = 0, \textit{bounded from one side on each end}\} \rangle$$

be the space spanned by the set of harmonic functions defined on M, which has the property that each one is bounded either from above or below on each end.

Observe that these spaces are monotonically contained in each other, i.e.,

$$\mathcal{H}_D(M) \subset \mathcal{H}_+(M) \subset \mathcal{H}'(M).$$

Let us also recalled the following potential theoretic definition.

Definition 4. *An end E of M is non-parabolic if it admits a positive Green's function with Neumann boundary condition on ∂E. Otherwise, E is said to be parabolic.*

It is important to note that if M has at least one non-parabolic end, then M admits a positive Green's function. In this case, we say that M is non-parabolic. The interested reader can refer to [11] for more detail descriptions. Let us now state the general theorem in [12].

Theorem (Li-Tam). *Let M be a complete, non-compact manifold without boundary. Then there exists a subspace $\mathcal{K} \subset \mathcal{H}'(M)$, such that, $\dim \mathcal{K}$ is equal to the number of ends of M.*

Moreover, if M is non-parabolic , then the subspace $\mathcal{K}$ can be taken to be in $\mathcal{H}_+(M)$. Also there exists another subspace $\mathcal{K}_N \subset \mathcal{H}_D(M)$, such that, $\dim \mathcal{K}_N$ is equal to the number of non-parabolic ends of M.

At this point, it is important to point out that even though an estimate on the dimension of the spaces $\mathcal{H}'(M)$, $\mathcal{H}_+(M)$, or $\mathcal{H}_D(M)$ will imply an estimate on the number of ends of corresponding type, however, in general, these spaces can be bigger than $\mathcal{K}$ or $\mathcal{K}_N$. Hence to effectively use the above theorem, one should use the constructive argument in the proof of the theorem to give an estimate on $\mathcal{K}$ and $\mathcal{K}_N$ directly. Indeed, this was the case in the proof of Theorem 1. This is also true for the two applications stated in the subsequence sections.

3. Manifolds with positive spectrum

A complete manifold (M, ds^2_M) is *conformally compact* if M is topologically a manifold with boundary given by ∂M. Moreover, there is a background metric ds^2_0 on $(M, \partial M)$ such that

$$ds^2_M = \rho^{-2}\, ds^2_0,$$

where ρ is a defining function for ∂M satisfying the conditions

$$\rho = 0 \quad \text{on} \quad \partial M$$

and

$$d\rho \neq 0 \quad \text{on} \quad \partial M.$$

A direct computation reveals that the sectional curvature, K_M, of the complete metric ds^2 has asymptotic value given by

$$K_M \sim -|d\rho|^2,$$

near ∂M. Hence if (M, ds^2_M) is also assumed to be Einstein with

$$Ric_M = -(m-1),$$

then

$$K_M(x) \sim -1,$$

as $x \to \infty$.

In 1999, Witten-Yau [19] proved a theorem concerning the AdS/CFT correspondence, which effectively ruled out the existence of worm holes. It is also a very interesting theorem in Riemannian geometry.

Theorem (Witten-Yau). *Let M^m be a conformally compact, Einstein manifold of dimension at least 3. Suppose the boundary ∂M of M has positive Yamabe constant, then*

$$H_{m-1}(M, \mathbb{Z}) = 0.$$

In particular, this implies that ∂M is connected and M must have only 1 end.

Shortly after, Cai-Galloway [2] relaxed the assumption of Witten-Yau by assuming the boundary ∂M has non-negative Yamabe constant. We would also like to point out that by a theorem of Schoen [17], a compact manifold has non-negative Yamabe constant is equivalent to the fact that it is conformally equivalent to a manifold with non-negative scalar curvature.

In his Stanford thesis, X. Wang [18] generalized the Witten-Yau, Cai-Galloway theorem by studying L^2 harmonic 1-forms.

Theorem (Wang). *Let M^m be a conformally compact manifold of dimension at least 3. Suppose the Ricci curvature of M is bounded by*

$$Ric_M \geq -(m-1)$$

and the lower bound of the spectrum of the Laplacian $\lambda_1(M)$ has a positive lower bound given by

$$\lambda_1(M) \geq (m-2),$$

then either

(1) *M has no non-constant L^2-harmonic 1-forms, i.e.,*

$$H^1(L^2(M)) = 0;$$

or

(2) *$M = \mathbb{R} \times N$ with the warped product metric*

$$ds_M^2 = dt^2 + \cosh^2 t \, ds_N^2,$$

where (N, ds_N^2) is a compact manifold with $Ric_N \geq -(m-2)$. Moreover, $\lambda_1(M) = m - 2$.

To see that this is indeed a generalization of the theorems of Witten-Yau and Cai-Galloway, one uses a theorem of Mazzeo [16] asserting that on a conformally compact manifold

$$H^1(L^2(M)) \simeq H^1(M, \partial M).$$

By a standard exact sequence argument, the conclusion that $H^1(L^2(M)) = 0$ implies that M has only 1 end. In addition to this, one also uses a theorem of Lee [10] giving a lower bound on λ_1 for conformally compact, Einstein manifold with non-negative Yamabe constant on ∂M.

Theorem (Lee). *Let M be a conformally compact, Einstein manifold with*

$$Ric_M = -(m-1).$$

Suppose that ∂M has non-negative Yamabe constant, then

$$\lambda_1(M) \geq \frac{(m-1)^2}{4}.$$

Since $\frac{(m-1)^2}{4} \geq m - 2$, Wang's theorem implies the theorems of Witten-Yau and Cai-Galloway. Observe that the warped product case in Wang's theorem has negative Yamabe constant on ∂M.

At this point, let us also recall a theorem of Cheng [5] stating that:

Theorem (Cheng). *Let M be a complete manifold with*

$$Ric_M \geq -(m-1),$$

then

$$\lambda_1(M) \leq \frac{(m-1)^2}{4}.$$

Combining the results of Cheng and Lee we conclude that

$$\lambda_1(M) = \frac{(m-1)^2}{4}$$

for conformally compact, Einstein manifolds, whose Ricci curvature is given by

$$Ric_M = -(m-1)$$

and has non-negative Yamabe constant for its boundary.

In the authors recent joint work with Jiaping Wang [14] , they proved this splitting type theorem without assuming the manifold is conformally compact.

Theorem 2 (Li-Wang). *Let M^m be a complete manifold with dimension $m \geq 3$. Suppose the Ricci curvature of M is bounded by*

$$Ric_M \geq -(m-1)$$

and

$$\lambda_1(M) \geq m-2,$$

then either

(1) *M has only 1 end with infinite volume;*

or

(2)*$M = \mathbb{R} \times N$ with the warped product metric*

$$ds_M^2 = dt^2 + \cosh^2 t\, ds_N^2,$$

where (N, ds_N^2) is compact with $Ric_N \geq -(m-2)$. Moreover, $\lambda_1(M) = m-2$.

It is worth noting that this theorem implies that when the lower bound for $\lambda_1(M)$ of Cheng is achieved, then either

(1)M has only 1 end with infinite volume,

or

(2)$M = \mathbb{R} \times N$ is the warped product and $m = 3$.

Also, since all the ends of a conformally compact manifold must have infinite volume, Theorem 2 is, in fact, a generalization of the theorems of Witten-Yau, Cai-Galloway, and Wang. It is also interesting to note that without the conformally compactness assumption, it is possible to have finite volume ends as indicated by following example.

Example 1. Let $M^m = \mathbb{R} \times N^{m-1}$ with the warped product metric

$$ds_M^2 = dt^2 + \exp(2t)\, ds_N^2,$$

where N is a compact manifold with

$$Ric_N \geq 0.$$

A direct computation shows that M has Ricci curvature bounded by

$$Ric_M \geq -(m-1)$$

and

$$\lambda_1(M) \geq m-2.$$

In fact, when $m = 3$, $\lambda_1(M) = 1$. Obviously M has two ends. One end E has infinite volume growth with

$$V_E(r) \sim C \exp((m-1)\, r),$$

while the other end e has finite volume with volume decay given by

$$V_e(\infty) - V_e(r) \sim C \exp(-(m-1)\, r).$$

We would like to point out that the pair of conditions

$$\left.\begin{aligned} Ric_M &\geq -(m-1) \\ \lambda_1(M) &\geq m-2 \end{aligned}\right\} \tag{1}$$

is equivalent to the pair of conditions

$$\left.\begin{aligned} Ric_M &\geq -\frac{m-1}{m-2}\lambda_1(M) \\ \lambda_1(M) &> 0. \end{aligned}\right\} \tag{2}$$

On the other hand, the pair of conditions

$$\left.\begin{aligned} Ric_M &\geq -\frac{m-1}{m-2}\lambda_1(M) \\ \lambda_1(M) &= 0 \end{aligned}\right\} \tag{3}$$

are equivalent to the single assumption that

$$Ric_M \geq 0,$$

because the condition $\lambda_1(M) = 0$ is a consequence of the curvature assumption.

Taking this point of view, Theorem 2 can be viewed as an analogue to the splitting theorem of Cheeger-Gromoll. Similarly to the fact that Theorem 1 is a perturbed version of the Cheeger-Gromoll splitting theorem, the following theorem in [14] is a perturbed version of Theorem 2.

Theorem 3 (Li-Wang). *Let M^m be a complete manifold with $m \geq 3$. Suppose $B_p(R) \subset M$ is a geodesic ball such that*

$$Ric_M \geq -(m-1) \qquad on \qquad M \setminus B_p(R)$$

and the lower bound of the spectrum of the Dirichlet Laplacian on $M \setminus B_p(R)$ is bounded by

$$\lambda_1(M \setminus B_p(R)) \geq m - 2 + \epsilon$$

for some $\epsilon > 0$. Then there exists a constant $0 < C(m, R, \alpha, v, \epsilon) < \infty$ depending only on m, R, $\alpha = \inf_{B_p(3R)} Ric_M$, $v = \inf_{x \in B_p(2R)} V_x(R)$, and ϵ, so that the number of infinite volume ends of M is at most $C(m, R, \alpha, v, \epsilon)$.

In both Theorem 2 and Theorem 3, the authors only managed to estimate the number of infinite volume ends by estimating the number of non-parabolic ends. In fact, when a manifold has positive spectrum, they proved that an end must either be non-parabolic with exponential volume growth, or it must be parabolic and finite volume with exponential volume decay. Moreover, these growth and decay estimates can be localized at each end.

Theorem 4 (Li-Wang). *Let M be a complete, non-compact, Riemannian manifold. Suppose E is an end of M given by a unbounded component of $M \setminus B_p(R)$, where $B_p(R)$ is a geodesic ball of radius R centered at some fixed point $p \in M$. Assume that the lower bound of the spectrum $\lambda_1(E)$ of the Dirichlet Laplacian on E is positive. Then as $r \to \infty$, either*

(1) *E is non-parabolic and has volume growth given by*

$$V_E(r) \geq C_1 \exp(2\sqrt{\lambda_1(E)}\, r)$$

for some constant $C_1 > 0$;
or

(2) *E is parabolic and has finite volume with volume decay given by*

$$V(E) - V_E(r) \leq C_2 \exp(-2\sqrt{\lambda_1(E)}\, r)$$

for some constant $C_2 > 0$.

In particular, if $\lambda_1(M) > 0$, then M must have exponential volume growth given by

$$V_p(r) \geq C_1 \exp(2\sqrt{\lambda_1(M)}\, r).$$

Both the volume growth and the volume decay estimates are sharp. For example, the growth estimate is achieved by the hyperbolic m-space, $\mathbb{H}^m$. Also, in Example 1 when dimension $m = 3$, the infinite volume end achieves the sharp volume growth estimate and the finite volume end achieves the sharp volume decay estimate. It is also interesting to point out that the sharp volume growth estimate is previously not known for manifolds with $\lambda_1(M) > 0$.

4. Minimal hypersurfaces

Let us recall that the well-known Bernstein's theorem (Bernstein, Fleming, Almgren, DeGiorgi, Simons) asserts that an entire minimal graph $M^m \subset \mathbb{R}^{m+1}$ must be linear if $m \leq 7$. Moreover, the dimension restriction is necessary as indicated by the examples of Bombieri, DeGiorgi, and Guisti. Since minimal graphs are necessarily area minimizing and hence stable (second variation of the area functional is non-negative), Fischer-Colbrie and Schoen [8] considered a generalization of Bernstein's theorem in this category. They proved that a complete, oriented, immersed, stable minimal surface in a complete manifold with non-negative scalar curvature must be conformally equivalent to either $\mathbb{C}$ or $\mathbb{R} \times \mathbb{S}^1$. Moreover, if the ambient manifold is $\mathbb{R}^3$ then the minimal surface must be planar. This special case was independently proved by do Carmo and Peng [6].

Later, Fischer-Colbrie [7] studied the structure of minimal surfaces with finite index. Recall that a minimal surface has finite index means that there are only a finite dimension of variations such that the second variations of the area functional is negative. In this case, Fischer-Colbrie proved that a complete, oriented, immersed, minimal surface with finite index in a complete manifold with non-negative

scalar curvature must be conformally equivalent to a compact Riemann surface with finitely many punctures. In particular, M must have finitely many ends. The special case when $N = \mathbb{R}^3$ was also independently proved by Gulliver [9]. It is in the spirit of the number of ends that Cao, Shen and Zhu [3] found a higher dimensional statement for stable minimal hypersurfaces in $\mathbb{R}^{m+1}$.

Theorem (Cao-Shen-Zhu). *Let $M^m \subset \mathbb{R}^{m+1}$ be a complete, oriented, immersed, stable minimal hypersurface in $\mathbb{R}^{m+1}$, then M must have only 1 end.*

This theorem is recently generalized to minimal hypersurfaces with finite index by the author and Jiaping Wang [13].

Theorem 5 (Li-Wang). *Let $M^m \subset \mathbb{R}^{m+1}$ be a complete, oriented, immersed, minimal hypersurface with finite index in $\mathbb{R}^{m+1}$, then M must have finitely many ends.*

In another paper [15], they also considered complete, properly immersed, stable (or with finite index) minimal hypersurfaces in a complete, non-negatively curved manifold.

Theorem 6 (Li-Wang). *Let $M^m \subset N^{m+1}$ be a complete, oriented, properly immersed, stable, minimal hypersurface. Suppose N is a complete manifold with non-negative sectional curvature. Then either*

(1) *M has only 1 end;*

or

(2) *$M = \mathbb{R} \times S$ with the product metric, where S is a compact manifold with non-negative sectional curvature. Moreover, M is totally geodesic in N.*

Theorem 7 (Li-Wang). *Let $M^m \subset N^{m+1}$ be a complete, oriented, properly immersed, minimal hypersurface with finite index. Suppose N is a complete manifold with non-negative sectional curvature. Then M must have finitely many ends.*

It is interesting to point out that in the case when $M = \mathbb{R} \times S$, the manifold is parabolic. In this case, it is necessary to estimate the space $\mathcal{K}$ rather than $\mathcal{K}'$. Again, the crucial point is to follow the construction of $\mathcal{K}$ and obtain sufficient estimates on the functions in $\mathcal{K}$ so that analytic techniques can be applied. In the case of Theorem 5, since the ambient manifold is $\mathbb{R}^{m+1}$ and hence the ends of M must all be non-parabolic, it is sufficient to estimate the space $\mathcal{K}'$ as stated in Theorem 2.

References

[1] M. Cai, Ends of Riemannian manifolds with nonnegative Ricci curvature outside a compact set, *Bull. AMS* **24** (1991), 371–377.

[2] M. Cai and G. J. Galloway, Boundaries of zero scalar curvature in the ADS/CFT correspondence, *Adv. Theor. Math. Phys.* **3** (1999), 1769–1783.

[3] H. Cao, Y. Shen, and S. Zhu, The structure of stable minimal hypersurfaces in $\mathbb{R}^{n+1}$, *Math. Res. Let.* **4** (1997), 637–644.

[4] J. Cheeger and D. Gromoll, The splitting theorem for manifolds of nonnegative Ricci curvature, *J. Diff. Geom.* **6** (1971), 119–128.

[5] S. Y. Cheng, Eigenvalue Comparison theorems and its Geometric Application, *Math. Z.* **143** (1975), 289–297.

[6] M. do Carmo and C. K. Peng, Stable complete minimal surfaces in $\mathbb{R}^3$ are planes, *Bull. AMS* **1** (1979), 903–906.

[7] D. Fischer-Colbrie, On complete minimal surfaces with finite Morse index in three manifolds, *Invent. Math.* **82** (1985), 121–132.

[8] D. Fischer-Colbrie and R. Schoen, The structure of complete stable minimal surfaces in 3-manifolds of non-negative scalar curvature, *Comm. Pure Appl. Math.* **33** (1980), 199–211.

[9] R. Gulliver, Index and total curvature of complete minimal surfaces., *Geometric measure theory and the calculus of variations (Arcata, Calif., 1984), Proc. Sympos. Pure Math. 44*, Amer. Math. Soc.,, Providence, RI., 1986, pp. 207–211.

[10] J. Lee, The spectrum of an asymptotic hyperbolic Einstein Manifold, *Comm. Anal. Geom.* **3** (1995), 253–271.

[11] P. Li, Curvature and function theory on Riemannian manifolds, *Survey in Differential Geometry "In Honor of Atiyah, Bott, Hirzebruch, and Singer"*, vol. VII, International Press, Cambridge, 2000, 71–111.

[12] P. Li and L. F. Tam, Harmonic functions and the structure of complete manifolds, *J. Diff. Geom.* **35** (1992), 359–383.

[13] P. Li and J. Wang, Minimal hypersurfaces with finite index, *Math. Res. Let.* **9** (2002), 95–103.

[14] P. Li and J. Wang, Complete manifolds with positive spectrum, *J. Diff. Geom.* **58** (2001), 501–534.

[15] P. Li and J. Wang, Stable minimal hypersurfaces in a nonnegatively curved manifold, *Preprint.*

[16] R. Mazzeo, The Hodge cohomology of a conformally compact metric, *J. Diff. Geom.* **28** (1988), 309–339.

[17] R. Schoen, Conformal deformation of a Riemannian metric to constant scalar curvature, *J. Diff. Geom.* **20** (1984), 479–495.

[18] X. Wang, On conformally compact Einstein manifolds, *Math. Res. Let.* **8** (2001), 671–688.

[19] E. Witten and S. T. Yau, Connectedness of the boundary in the AdS/CFT correspondence, *Adv. Theor. Math. Phys.* **3** (1999), 1635–1655.

ICM 2002 · Vol. II · 303–313

Index Iteration Theory for Symplectic Paths with Applications to Nonlinear Hamiltonian Systems

Yiming Long*

Abstract

In recent years, we have established the iteration theory of the index for symplectic matrix paths and applied it to periodic solution problems of nonlinear Hamiltonian systems. This paper is a survey on these results.

2000 Mathematics Subject Classification: 58E05, 70H05, 34C25.
Keywords and Phrases: Iteration theory, Index, Symplectic path, Hamiltonian system, Periodic orbit.

Since P. Rabinowitz's pioneering work [35] of 1978, variational methods have been widely used in the study of existence of solutions of Hamiltonian systems. But how to study the geometric multiplicity and stability of periodic solution orbits obtained by variational methods has kept to be a difficulty problem. For example let $x = x(t)$ be a τ-periodic solution of a Hamiltonian system

$$\dot{x}(t) = JH'(x(t)), \quad \forall t \in \mathbf{R}. \tag{0.1}$$

The m-th iteration x^m of x is defined by induction $m-1$ times via $x(t+\tau) = x(t)$ for $t > 0$. It runs m-times along the orbit of x. Geometrically these iterations produce the same solution orbit of (0.1), but they are different as critical points of corresponding functionals. This multiple covering phenomenon causes major difficulties in the study.

A natural way to study solution orbits found by variational methods is to study the Morse-type index sequences of their iterations. But when one studies general Hamiltonian systems, the Morse indices of the critical points of the corresponding functional are always infinite. To overcome this difficulty, in their celebrated paper [6] of 1984, C. Conley and E. Zehnder defined an index theory for any nondegenerate paths in $\mathrm{Sp}(2n)$ with $n \geq 2$, i.e., the so called Conley-Zehnder index

*Nankai Institute of Mathematics, Nankai University, Tianjin 300071, China. E-mail: longym@nankai.edu.cn

theory. This index theory was further defined for non-degenerate paths in Sp(2) by E. Zehnder and the author in [33] of 1990. The index theory for degenerate linear Hamiltonian systems was defined by C. Viterbo in [39] and the author in [20] of 1990 independently. In [25] of 1997, this index was extended to any degenerate symplectic matrix paths.

Motivated by the iteration theories for the Morse type index theories established by R. Bott in 1956 and by I. Ekeland in 1980s, in recent years the author extended the index theory mentioned above, introduced an index function theory for symplectic matrix paths, and established the iteration theory for the index theory of symplectic paths. Applying this index iteration theory to nonlinear Hamiltonian systems, interesting results on periodic solution problems of Hamiltonian systems are obtained. Here a brief survey is given on these subjects. Readers are referred to the author's recent book [30] for further details.

1. Index function theory for symplectic paths

As usual we define the symplectic group by $\mathrm{Sp}(2n) = \{M \in \mathrm{GL}(\mathbf{R}^{2n}) \,|\, M^T J M = J\}$, where $J = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$, I is the identity matrix on $\mathbf{R}^n$, and M^T denotes the transpose of M. For $\omega \in \mathbf{U}$, the unit circle in the complex plane $\mathbf{C}$, we define the ω-singular subset in $\mathrm{Sp}(2n)$ by $\mathrm{Sp}(2n)^0_\omega = \{M \in \mathrm{Sp}(2n) \,|\, \omega^{-n} \det(\gamma(\tau) - \omega I) = 0\}$. Here for any $M \in \mathrm{Sp}(2n)^0_\omega$, we define the orientation of $\mathrm{Sp}(2n)^0_\omega$ at M by the positive direction $\frac{d}{dt} M \exp(tJ)|_{t=0}$. Since the fundamental solution of a general linear Hamiltonian system with continuous symmetric periodic coefficient $2n \times 2n$ matrix function $B(t)$,

$$\dot{x}(t) = JB(t)x(t), \quad \forall t \in \mathbf{R}, \tag{1.1}$$

is a path in $\mathrm{Sp}(2n)$ starting from the identity, for $\tau > 0$ we define the set of symplectic matrix paths by $\mathcal{P}_\tau(2n) = \{\gamma \in C([0,\tau], \mathrm{Sp}(2n)) \,|\, \gamma(0) = I\}$. For any two path ξ and $\eta : [0,\tau] \to \mathrm{Sp}(2n)$ with $\xi(\tau) = \eta(0)$, as usual we define $\eta * \xi(t)$ by $\xi(2t)$ if $0 \le t \le \tau/2$, and $\eta(2t - \tau)$ if $\tau/2 \le t \le \tau$. We define a special path $\zeta : [0,\tau] \to \mathrm{Sp}(2n)$ by

$$\zeta(t) = \mathrm{diag}(2 - \frac{t}{\tau}, \ldots, 2 - \frac{t}{\tau}, (2 - \frac{t}{\tau})^{-1}, \ldots, (2 - \frac{t}{\tau})^{-1}), \quad \text{for } 0 \le t \le \tau.$$

Definition 1. (cf. [27]) *For any $\tau > 0$, $\omega \in \mathbf{U}$, and $\gamma \in \mathcal{P}_\tau(2n)$, we define the ω-nullity of γ by*

$$\nu_\omega(\gamma) = \dim_{\mathbf{C}} \ker_{\mathbf{C}}(\gamma(\tau) - \omega I). \tag{1.2}$$

If γ is ω non-degenerate, i.e., $\nu_\omega(\gamma) = 0$, we define the ω-index of γ by the intersection number

$$i_\omega(\gamma) = [\mathrm{Sp}(2n)^0_\omega : \gamma * \zeta]. \tag{1.3}$$

If γ is ω degenerate, i.e., $\nu_\omega(\gamma) > 0$, we let $\mathcal{F}(\gamma)$ be the set of all open neighborhoods of γ in $\mathcal{P}_\tau(2n)$, and define

$$i_\omega(\gamma) = \sup_{U \in \mathcal{F}(\gamma)} \inf\{i_\omega(\beta) \,|\, \beta \in U, \nu_\omega(\beta) = 0\}. \tag{1.4}$$

Then we call $(i_\omega(\gamma), \nu_\omega(\gamma)) \in \mathbf{Z} \times \{0, 1, \ldots, 2n\}$ *the index function of* γ *at* ω.

The relation of this index $(i_1(\gamma), \nu_1(\gamma))$ with the Morse index of τ-periodic solutions of the problem (1) was proved by C. Conley, E. Zehnder, and the author in [6], [33], and [20] (cf. Theorem 6.1.1 of [30]).

2. Iteration theory of the index for symplectic paths

Given a path $\gamma \in \mathcal{P}_\tau(2n)$, its iteration is defined inductively by $\gamma(t+\tau) = \gamma(t)\gamma(\tau)$ for $t \geq 0$, i.e.,

$$\gamma^m(t) = \gamma(t - j\tau)\gamma(\tau)^j, \qquad j\tau \leq t \leq (j+1)\tau, j = 0, 1, \ldots, m-1, \tag{2.1}$$

for any m in the natural integer set $\mathbf{N}$. For our applications of this index theory to nonlinear Hamiltonian systems, we are facing two types of problems:

$\langle 1 \rangle$ knowing the end point $\gamma(\tau)$ of a path $\gamma \in \mathcal{P}_\tau(2n)$, the initial index $(i_1(\gamma), \nu_1(\gamma))$, and the iteration time m, want to find the index $i_1(\gamma^m)$ of the m-th iterated path γ^m;

$\langle 2 \rangle$ knowing the end point $\gamma(\tau)$ of a path $\gamma \in \mathcal{P}_\tau(2n)$, the initial index $(i_1(\gamma), \nu_1(\gamma))$, and the index $(i_1(\gamma^m), \nu_1(\gamma^m))$ of the m-th iterated path γ^m, want to find the iteration time m.

To solve these problems, we first generalize Bott's formula of the iterated Morse index for closed geodesics to the index theory for general symplectic paths:

Theorem 2 (cf. [27]). *For any* $\tau > 0$, $\gamma \in \mathcal{P}_\tau(2n)$, $z \in \mathbf{U}$, *and* $m \in \mathbf{N}$, *there hold:*

$$i_z(\gamma^m) = \sum_{\omega^m = z} i_\omega(\gamma), \quad \nu_z(\gamma^m) = \sum_{\omega^m = z} \nu_\omega(\gamma). \tag{2.2}$$

By (2.2) it is easy to see that the mean index $\hat{i}(\gamma) = \lim_{m \to +\infty} i_1(\gamma^m)/m$ for any $\gamma \in \mathcal{P}_\tau(2n)$ is always a finite real number.

To further solve the problems $\langle 1 \rangle$ and $\langle 2 \rangle$, we need to go beyond the Bott-type formula (2.2). For a given path γ we consider to deform it to a new path η in $\mathcal{P}_\tau(2n)$ so that

$$i_1(\gamma^m) = i_1(\eta^m), \quad \nu_1(\gamma^m) = \nu_1(\eta^m), \quad \forall m \in \mathbf{N}, \tag{2.3}$$

and that $(i_1(\eta^m), \nu_1(\eta^m))$ is easy enough to compute. This leads to finding homotopies $\delta : [0,1] \times [0,\tau] \to \mathrm{Sp}(2n)$ starting from γ in $\mathcal{P}_\tau(2n)$ and keeping the end points of the homotopy always stay in a certain suitably chosen maximal subset of $\mathrm{Sp}(2n)$ so that (2.3) always holds. By (2.2), this set is defined to be the path connected component $\Omega^0(M)$ containing $M = \gamma(\tau)$ of the set

$$\begin{aligned} \Omega(M) = \{N \in \mathrm{Sp}(2n) \quad | \quad & \sigma(N) \cap \mathbf{U} = \sigma(M) \cap \mathbf{U}, \text{ and} \\ & \nu_\lambda(N) = \nu_\lambda(M)\ \forall\, \lambda \in \sigma(M) \cap \mathbf{U}\}. \end{aligned} \tag{2.4}$$

Here we call $\Omega^0(M)$ the *homotopy component* of M in $\mathrm{Sp}(2n)$.

Using normal forms of symplectic matrices (cf. [32], [13]), we then decompose $\gamma(\tau)$ within $\Omega^0(\gamma(\tau))$ into product of 10 special 2×2 and 4×4 symplectic normal

form matrices, which we call *basic normal forms*. Correspondingly by the homotopy invariance and symplectic additivity of the index theory, the computations in (2.3) are reduced to iterations of those paths in Sp(2) or Sp(4) whose end points are one of the 10 basic normal form matrices. The study of the index for iterations of any symplectic paths is carried out for paths in Sp(2) via the $\mathbf{R}^3$-cylindrical coordinate representation of Sp(2), then for hyperbolic and elliptic paths in Sp($2n$). This yields the precise iteration formula obtained in [29] of the index theory for any symplectic path $\gamma \in \mathcal{P}_\tau(2n)$ in terms of the basic norm form decomposition of $\gamma(\tau)$, $(i(\gamma,1),\nu(\gamma,1))$, and the iteration time m.

For any $M \in \mathrm{Sp}(2n)$, its splitting numbers at an $\omega \in \mathbf{U}$ is defined in [27] by

$$S_M^{\pm}(\omega) = \lim_{\epsilon \to 0^+} i_{\omega \exp(\pm\sqrt{-1}\epsilon)}(\gamma) - i_\omega(\gamma), \tag{2.5}$$

via any $\gamma \in \mathcal{P}_\tau(2n)$ satisfying $\gamma(\tau) = M$. Then it is proved that the splitting numbers of M at ω can be characterized algebraically.

Motivated by the precise iteration formulae of [29], the following second index iteration formula of any symplectic path is established by C. Zhu and the author. Here we denote by $(i(\gamma,m),\nu(\gamma,m)) = (i_1(\gamma^m),\nu_1(\gamma^m))$.

Theorem 3 (cf. [34]). *For any $\tau > 0$, $\gamma \in \mathcal{P}_\tau(2n)$, and $m \in \mathbf{N}$, there holds:*

$$\begin{aligned} i(\gamma,m) &= m(i(\gamma,1) + S_M^+(1) - C(M)) \\ &\quad +2 \sum_{\theta \in (0,2\pi)} E(\frac{m\theta}{2\pi}) S_M^-(e^{\sqrt{-1}\theta}) - (S_M^+(1) + C(M)), \end{aligned} \tag{2.6}$$

where $M = \gamma(\tau)$, $C(M) = \sum_{0<\theta<2\pi} S_M^-(e^{\sqrt{-1}\theta})$, and $E(a) = \min\{k \in \mathbf{Z} \,|\, k \geq a\}$ for any $a \in \mathbf{R}$.

In order to solve problems on nonlinear Hamiltonian systems, various index iteration inequalities for any path $\gamma \in \mathcal{P}_\tau(2n)$ and $m \in \mathbf{N}$ are proved by D. Dong, C. Liu, C. Zhu and the author in [7], [16], [17], and [34].

Theorem 4. *For any $\gamma \in \mathcal{P}_\tau(2n)$ and $m \in \mathbf{N}$, the following iteration inequalities always hold.*

Estimate via mean index (cf. [16], [17]):

$$m\hat{i}(\gamma) - n \leq i(\gamma,m) \leq m\hat{i}(\gamma) + n - \nu(\gamma,m). \tag{2.7}$$

Estimate via initial index (cf. [18]):

$$\begin{aligned} m(i(\gamma,1) &+ \nu(\gamma,1) - n) + n - \nu(\gamma,1) \leq i(\gamma,m) \\ &\leq m(i(\gamma,1) + n) - n - (\nu(\gamma,m) - \nu(\gamma,1)). \end{aligned} \tag{2.8}$$

Successive index estimate (cf. [34]):

$$\begin{aligned} \nu(\gamma,m) - \frac{e(\gamma(\tau))}{2} &\leq i(\gamma,m+1) - i(\gamma,m) - i(\gamma,1) \\ &\leq \nu(\gamma,1) - \nu(\gamma,m+1) + \frac{e(\gamma(\tau))}{2}. \end{aligned} \tag{2.9}$$

Here we define $e(M)$ to be the total multiplicity of eigenvalues of M on $\mathbf{U}$ and call it the elliptic height of M.

A consequence of the iteration inequality (2.8) together with the necessary and sufficient conditions for any equality in (2.8) to hold for some m yields a new proof of the following theorem of D. Dong and the author on controlling the iteration time m via indices:

Theorem 5 (cf. [7]). *For any $\gamma \in \mathcal{P}_\tau(2n)$ and $m \in \mathbf{N}$, suppose $i(\gamma, m) \leq n+1$, $i(\gamma, 1) \geq n$, and $\nu(\gamma, 1) \geq 1$. Then $m = 1$.*

Note also that the inequality (2.9) yields a way to estimate the ellipticity of solution orbits of Hamiltonian systems obtained by variational methods via their iterated indices.

In order to study the properties of solution orbits of the system (0.1) on a given energy hypersurface, when the number of orbits is finite, we need to study common properties of any given finite family of symplectic paths $\gamma_j \in \mathcal{P}_{\tau_j}(2n)$ with $1 \leq j \leq q$. This leads to the following common index jump theorem of C. Zhu and the author proved in [34]. For any $\gamma \in \mathcal{P}_\tau(2n)$, its m-th index jump $\mathcal{G}_m(\gamma)$ is defined to be the open interval $\mathcal{G}_m(\gamma) = (i(\gamma, m) + \nu(\gamma, m) - 1, i(\gamma, m+2))$.

Theorem 6 (cf. [34]). *Let $\gamma_j \in \mathcal{P}_{\tau_j}(2n)$ with $1 \leq j \leq q$ satisfying*

$$\hat{i}(\gamma_j) > 0, \quad i(\gamma_j, 1) \geq n, \qquad 1 \leq j \leq q. \tag{2.10}$$

Then there exist infinitely many positive integer tuples $(N, m_1, \ldots, m_q) \in \mathbf{N}^{q+1}$ such that

$$\emptyset \neq [2N - \kappa_1, 2N + \kappa_2] \subset \bigcap_{j=1}^{q} \mathcal{G}_{2m_j - 1}(\gamma_j), \tag{2.11}$$

where $\kappa_1 = \min_{1 \leq j \leq q}(i(\gamma_j, 1) + 2S^+_{\gamma_j(\tau_j)}(1) - \nu(\gamma_j, 1))$ and $\kappa_2 = \min_{1 \leq j \leq q} i(\gamma_j, 1) - 1$.

In order to prove this theorem, we need to make each index jump to be as big as possible, and to make their largest sizes happen simultaneously to guarantee the existence of a non-empty largest common intersection interval among them. By the term $E(\frac{m\theta}{2\pi})$ in the abstract iteration formula (2.6), such a problem is reduced to a dynamical system problem on a torus, and is solved by properties of closed additive subgroups of tori.

3. Applications to nonlinear Hamiltonian systems

So far, we have applied our index iteration theory to three important problems on periodic solutions of nonlinear Hamiltonian systems. Let $T > 0$ and suppose x is a non-constant T-periodic solution of the nonlinear Hamiltonian system (0.1). Suppose the minimal period of x is $\tau = T/k$ for some $k \in \mathbf{N}$. We denote by $\gamma_x \in \mathcal{P}_\tau(2n)$ the fundamental solution of the linearized Hamiltonian system (1.1) at x with $B(t) = H''(x(t))$, and the iterated index of x by $(i(x, m), \nu(x, m)) = (i(\gamma_x, m), \nu(\gamma_x, m))$ for all $m \in \mathbf{N}$.

3.1. Prescribed minimal period solution problem

In [35] of 1978, P. Rabinowitz posed a conjecture on whether the Hamiltonian system possesses periodic solutions with prescribed minimal period when the Hamiltonian function satisfies his superquadratic conditions. This conjecture is studied by D. Dong and the author as an application of our index iteration theory. Note that for a non-constant τ-periodic solution x of the autonomous system (0.1), the condition on the nullity in Theorem 5 always holds. Thus Theorem 5 yields:

Theorem 7 (cf. [7]). *For any non-constant τ-periodic solution x of (0.1), denote its minimal period by τ/m for some $m \in \mathbf{N}$. Suppose $i(x|_{[0,\tau]}, 1) \leq n+1$ and $n \leq i(x|_{[0,\tau/m]}, 1)$. Then $m = 1$, i.e., τ is the minimal period of x.*

Here the first estimate on the index holds if x is obtained by minimax or minimization methods, and the second estimate on the index holds if the Hamiltonian function H is convex in a certain weak sense along the orbit of x. This result reveals the intrinsic relationship between the minimal period of a periodic solution and its indices, and unifies all the results on Rabinowitz's conjecture under various convexity conditions. Specially, it recovers the famous theorem of I. Ekeland and H. Hofer in 1985 (cf. [11]) who solved Rabinowitz's conjecture for convex superquadratic Hamiltonian systems.

3.2. Periodic points of the Poincaré map of Lagrangian systems on tori

In 1984, C. Conley stated a conjecture on whether the Poincaré map of any 1-periodic time dependent Hamiltonian system defined on the standard torus $T^{2n} = \mathbf{R}^{2n}/\mathbf{Z}^{2n}$ always possesses infinitely many periodic points which are produced by contractible periodic solutions of the corresponding Hamiltonian system on T^{2n}. A celebrated partial answer to this conjecture was given by D. Salamon and E. Zehnder in 1992 (cf. [37]) for a large class of symplectic manifolds on which every contractible integer periodic solution of the Hamiltonian system has at least one Floquet multiplier not equal to 1. So far Conley conjecture is still open and seems far from being completely understood.

In [28], we studied the Lagrangian system version of this conjecture. Consider

$$\frac{d}{dt}L_{\dot{x}}(t,x,\dot{x}) - L_x(t,x,\dot{x}) = 0, \quad x \in \mathbf{R}^n, \tag{3.1}$$

where $L_{\dot{x}}$ and L_x denote the gradients of L with respect to $\dot{x}$ and x respectively. The main result is the following:

Theorem 8 (cf. [28]). *Suppose the Lagrangian function L satisfies*

(L1) $L(t,x,p) = \frac{1}{2}A(t)p \cdot p + V(t,x)$, where $\frac{1}{2}A(t)p \cdot p \geq \lambda|p|^2$ for all $(t,p) \in \mathbf{R} \times \mathbf{R}^n$ and some fixed constant $\lambda > 0$.

(L2) $A \in C^3(\mathbf{R}, \mathcal{L}_s(\mathbf{R}^n))$, $V \in C^3(\mathbf{R} \times \mathbf{R}^n, \mathbf{R})$, both A and V are 1-periodic in all of their variables, where $\mathcal{L}_s(\mathbf{R}^n)$ denotes the set of $n \times n$ real symmetric matrices.

Then the Poincaré map Ψ of the system (3.1) possesses infinitely many periodic points on TT^n produced by contractible integer periodic solutions of the system (3.1).

In the proof of Theorem 8, the above inequality (2.7) plays a crucial role. By this inequality, at very high iteration level, a global homological injection map can be constructed which maps a generator of a certain non-trivial local critical group to a nontrivial homology class $[\sigma]$ in a global homology group, if the number of contractible integer periodic solution towers of the system (3.1) is finite. But on the other hand, by a technique of V. Bangert and W. Klingenberg in [3], it is shown that this homology class $[\sigma]$ must be trivial globally. This contradiction then yields the conclusion of Theorem 8.

3.3. Closed characteristics on convex compact hypersurfaces

Denote the set of all compact strictly convex C^2-hypersurfaces in $\mathbf{R}^{2n}$ by $\mathcal{H}(2n)$. For $\Sigma \in \mathcal{H}(2n)$ and $x \in \Sigma$, let $N_\Sigma(x)$ be the outward normal unit vector at x of Σ. We consider the problem of finding $\tau > 0$ and a curve $x \in C^1([0,\tau], \mathbf{R}^{2n})$ such that

$$\begin{cases} \dot{x}(t) &= JN_\Sigma(x(t)), \quad x(t) \in \Sigma, \quad \forall t \in \mathbf{R}, \\ x(\tau) &= x(0). \end{cases} \tag{3.1}$$

A solution (τ, x) of the problem (3.1) is called a *closed characteristic* on Σ. Two closed characteristics (τ, x) and (σ, y) are *geometrically distinct*, if $x(\mathbf{R}) \neq y(\mathbf{R})$. We denote by $\mathcal{T}(\Sigma)$ the set of all geometrically distinct closed characteristics (τ, x) on Σ with τ being the minimal period of x. Note that the problem (3.1) can be described in a Hamiltonian system version and solved by variational methods. A closed characteristic (τ, x) is *non-degenerate*, if 1 is a Floquet multiplier of x of precisely algebraic multiplicity 2, and is *elliptic*, if all the Floquet multipliers of x are on $\mathbf{U}$. Let $^\#A$ denote the total number of elements in a set A.

This problem has been studied for more than 100 years since at least A. M. Liapunov in 1892. A long standing conjecture on the multiplicity of closed characteristics is whether

$$^\#\tilde{\mathcal{J}}(\Sigma) \geq n, \qquad \forall\, \Sigma \in \mathcal{H}(2n). \tag{3.2}$$

The first break through on this problem in the global sense was made by P. Rabinowitz [35] and A. Weinstein [40] in 1978. They proved $^\#\mathcal{T}(\Sigma) \geq 1$ for all $\Sigma \in \mathcal{H}(2n)$. Besides many results under pinching conditions, in 1987–1988, I. Ekeland-L. Lassoued, I. Ekeland-H. Hofer, and A, Szulkin proved $^\#\mathcal{T}(\Sigma) \geq 2$ for all $\Sigma \in \mathcal{H}(2n)$ and $n \geq 2$. In 1998, H. Hofer, K. Wysocki, and E. Zehnder proved in [14]: $^\#\mathcal{T}(\Sigma) = 2$ or $+\infty$ for every $\Sigma \in \mathcal{H}(4)$. In recent years C. Liu, C. Zhu, and the author gave the following answers to the conjecture (3.2):

Theorem 9 (cf. [34]). *There holds*

$$^\#\mathcal{T}(\Sigma) \geq [\frac{n}{2}] + 1, \quad \forall \Sigma \in \mathcal{H}(2n), \tag{3.3}$$

where $[a] = \max\{k \in \mathbf{Z} \,|\, k \leq a\}$ *for any* $a \in \mathbf{R}$. *Moreover, if all the closed characteristics on* Σ *are non-degenerate, then* $^\#\mathcal{T}(\Sigma) \geq n$.

Theorem 10 (cf. [19]). *For any* $\Sigma \in \mathcal{H}(2n)$, *if* Σ *is symmetric with respect to the origin, i.e.,* $x \in \Sigma$ *implies* $-x \in \Sigma$, *then* $^\#\mathcal{T}(\Sigma) \geq n$.

Very recently, Y. Dong and the author further proved the following result.

Theorem 11 (cf. [8]). *Let $\Sigma \in \mathcal{H}(2n)$ be P-symmetric with respect to the origin, i.e., $x \in \Sigma$ implies $Px \in \Sigma$, where $P = \mathrm{diag}(-I_{n-k}, I_k, -I_{n-k}, I_k)$ for some fixed integer $k \in [0, n-1]$. Let $\Sigma(k) = \{(x,y) \in (\mathbf{R}^k)^2 \,|\, (0,x,0,y) \in \Sigma\}$. Suppose ${}^{\#}\mathcal{T}(\Sigma(k)) \leq k$ or ${}^{\#}\mathcal{T}(\Sigma(k)) = +\infty$ holds. Then ${}^{\#}\mathcal{T}(\Sigma) \geq n - 2k$.*

Proof of Theorem 11 depends on a new index iteration theory for symplectic paths iterated by the formula $\gamma(t+\tau) = P\gamma(t)P\gamma(\tau)$ for $t \geq 0$.

The second long standing conjecture on closed characteristics is whether there always exists at least an elliptic closed characteristic on any $\Sigma \in \mathcal{H}(2n)$. Up to the author's knowledge, the existence of one elliptic closed characteristic on $\Sigma \in \mathcal{H}(2n)$ was proved by I. Ekeland in 1990 when Σ is $\sqrt{2}$-pinched by two spheres, and by G.-F. Dell'Antonio, B. D'Onofrio, and I. Ekeland in 1992 when Σ is symmetric with respect to the origin. Recently using an enhanced version of the iteration estimate (2.9) on the elliptic height, based on results in [29] the following result was further proved by C. Zhu and the author.

Theorem 12 (cf. [34]). *For $\Sigma \in \mathcal{H}(2n)$, suppose ${}^{\#}\mathcal{T}(\Sigma) < +\infty$. Then there exists at least an elliptic closed characteristic on Σ. Moreover, suppose $n \geq 2$ and ${}^{\#}\mathcal{T}(\Sigma) \leq 2[n/2]$. Then there exist at least two elliptic elements in $\mathcal{T}(\Sigma)$.*

The main ingredient in the proofs of Theorems 9 to 12 is our index iteration theory mentioned above. To illustrate this method, we briefly describe below the main idea in the proof of (3.3) in Theorem 9. Because each closed characteristic on Σ corresponds to infinitely many critical values of the related dual action functional, our way to solve the problem is to study how the index intervals of iterated closed characteristics cover the set of integers $2\mathbf{N} - 2 + n$ to count the number of closed characteristics on Σ. Suppose $q = {}^{\#}\tilde{\mathcal{J}}(\Sigma) < +\infty$. In the proof of the multiplicity claim (3.3) of Theorem 9, the most important ingredient is the following estimates:

$$\begin{aligned} q &\geq {}^{\#}\left((2\mathbf{N} - 2 + n) \cap \cap_{j=1}^{q} \mathcal{G}_{2m_j-1}(\gamma_{x_j})\right) \\ &\geq {}^{\#}\left((2\mathbf{N} - 2 + n) \cap [2N - \kappa_1, 2N + \kappa_2]\right) \\ &\geq [\frac{n}{2}] + 1, \end{aligned} \tag{3.4}$$

The first inequality in (3.4) is a new version of the Liusternik-Schnirelman theoretical argument at the iterated index level, which distinguishes solution orbits geometrically instead of critical points only as usual methods do. The second inequality in (3.4) uses the common index jump Theorem 6. The last inequality in (3.4) uses the Morse theoretical approach. Roughly speaking, the common index jump theorem picks up as many as possible points of $2\mathbf{N} - 2 + n$ in the interval $[2N - \kappa_1, 2N + \kappa_2] \subset \cap_{j=1}^{q} \mathcal{G}_{2m_j-1}(\gamma_{x_j})$, which yields a lower bound for ${}^{\#}\mathcal{T}(\Sigma)$.

As usual, a hypersurface $\Sigma \subset \mathbf{R}^{2n}$ is star-shaped if the tangent hyperplane at any $x \in \Sigma$ does not intersect the origin. Closed characteristics on Σ can be defined by (3.1) too. In this case, the result ${}^{\#}\tilde{\mathcal{J}}(\Sigma) \geq 1$ was proved by P. Rabinowitz in [35] of 1978. Then multiplicity results were proved under certain pinching conditions on star-shaped Σ. Recently, the following result for the free case was proved by X. Hu and the author:

Theorem 13 (cf. [15]). *Let Σ be a star-shaped compact C^2-hypersurface in $\mathbf{R}^{2n}$. Suppose all the closed characteristics on Σ and all of their iterates are non-*

degenerate. Then ${}^{\#}\mathcal{T}(\Sigma) \geq 2$. *Moreover, if* $n = 2$ *and* ${}^{\#}\mathcal{T}(\Sigma) < +\infty$ *further holds, then there exist at least two elliptic closed characteristics on* Σ.

Here the crucial point is to prove $i(x,1) \geq n$ when (τ, x) is the only geometrically distinct closed characteristic on Σ. This conclusion is proved by using our index iteration theory and an identity of non-degenerate closed characteristics on Σ proved by C. Viterbo in 1989.

Because of Theorem 9 and other indications, we suspect that the following holds:

$$\{{}^{\#}\mathcal{T}(\Sigma) \,|\, \Sigma \in \mathcal{H}(2n)\} = \{k \in \mathbf{Z} \,|\, [\frac{n}{2}] + 1 \leq k \leq n\} \cup \{+\infty\}. \tag{3.5}$$

We also suspect that closed orbits of the Reeb field on a compact contact hypersurfaces in a symplectic manifold may have similar properties.

Many other problems related to iterations of periodic solution orbits are still open, for example, the Seifert conjecture on the existence of at least n brake orbits for the given energy problem of classical Hamiltonian systems on $\mathbf{R}^n$ (cf. [38], [1] and the references there in), and the conjecture on the existence of infinitely many geometrically distinct closed geodesics on every compact Riemannian manifold (cf. [2] and the solution for S^2 by J. Franks and V. Bangert). We believe that our index iteration theory for symplectic paths and the methods we developed to establish and apply it to nonlinear problems will have the potential to play more roles in the study on these problems and in other mathematical areas.

Acknowledgements. The author sincerely thanks the 973 Program of MOST, NNSF, MCME, RFDP, PMC Key Lab of MOE of China, S. S. Chern Foundation, CEC of Tianjin, and Qiu Shi Sci. Tech. Foundation of Hong Kong for their supports in recent years.

References

[1] A. Ambrosetti, V. Benci, & Y. Long, A note on the existence of multiple brake orbits. *Nonlinear Anal. TMA.* 21 (1993), 643–649.

[2] V. Bangert, Geodetische Linien auf Riemannschen Mannigfaltigkeiten. *Jber. D. Dt. Math.-Verein.* 87 (1985), 39–66.

[3] V. Bangert & W. Klingenberg, Homology generated by iterated closed geodesics. *Topology.* 22 (1983), 379–388.

[4] R. Bott, On the iteration of closed geodesics and the Sturm intersection theory. *Comm. Pure Appl. Math.* 9 (1956), 171–206.

[5] K. C. Chang, Infinite Dimensional Morse Theory and Multiple Solution Problems. Birkhäuser. Basel. (1993).

[6] C. Conley & E. Zehnder, Morse-type index theory for flows and periodic solutions for Hamiltonian equations. *Comm. Pure Appl. Math.* 37 (1984), 207–253.

[7] D. Dong & Y. Long, The iteration formula of the Maslov-type index theory with applications to nonlinear Hamiltonian systems. *Trans. Amer. Math. Soc.* 349 (1997), 2619–2661.

[8] Y. Dong & Y. Long, Closed characteristics on partially symmetric convex hypersurfaces in $\mathbf{R}^{2n}$. (2002) Preprint.

[9] I. Ekeland, Une théorie de Morse pour les systémes hamiltoniens convexes. *Ann. IHP. Anal. non Linéaire.* 1 (1984), 19–78.
[10] I. Ekeland, Convexity Methods in Hamiltonian Mechanics. Springer. Berlin. 1990.
[11] I. Ekeland & H. Hofer, Periodic solutions with prescribed period for convex autonomous Hamiltonian systems. *Invent. Math.* 81 (1985), 155–188.
[12] I. Ekeland & H. Hofer, Convex Hamiltonian energy surfaces and their closed trajectories. *Comm. Math. Phys.* 113 (1987), 419–467.
[13] J. Han & Y. Long, Normal forms of symplectic matrices (II). *Acta Sci. Nat. Univ. Nankai.* 32 (1999) 30–41.
[14] H. Hofer, K. Wysocki, & E. Zehnder, The dynamics on three-dimensional strictly convex energy surfaces. *Ann. of Math.* 148 (1998) 197–289.
[15] X. Hu & Y. Long, Multiplicity of closed characteristics on non-degenerate star-shaped hypersurfaces in $\mathbf{R}^{2n}$. *Sciences in China* (2002) to appear.
[16] C. Liu & Y. Long, An optimal increasing estimate for iterated Maslov-type indices. *Chinese Sci. Bull.* 42 (1997), 2275–2277.
[17] C. Liu & Y. Long, Iteration inequalities of the Maslov-type index theory with applications. *J. Diff. Equa.* 165 (2000) 355–376.
[18] C. Liu & Y. Long, Iterated index formulae for closed geodesics with applications. *Science in China.* 45 (2002) 9–28.
[19] C. Liu, Y. Long, & C. Zhu, Multiplicity of closed characteristics on symmetric convex hypersurfaces in $\mathbf{R}^{2n}$. *Math. Ann.* (to appear).
[20] Y. Long, Maslov-type index, degenerate critical points, and asymptotically linear Hamiltonian systems. *Science in China (Scientia Sinica).* Series A. 7 (1990), 673–682. (Chinese edition), 33 (1990), 1409–1419. (English edition).
[21] Y. Long, Index Theory of Hamiltonian Systems with Applications. Science Press. Beijing. 1993. (In Chinese).
[22] Y. Long, The minimal period problem for classical Hamiltonian systems with even potentials. *Ann. Inst. H. Poincare. Anal. non linéaire.* 10 (1993), 605–626.
[23] Y. Long, The minimal period problem of periodic solutions for autonomous superquadratic second order Hamiltonian systems. *J. Diff. Equa.* 111 (1994), 147–174.
[24] Y. Long, On the minimal period for periodic solutions of nonlinear Hamiltonian systems. *Chinese Ann. of Math.* 18B (1997), 481–484.
[25] Y. Long, A Maslov-type index theory for symplectic paths. *Top. Meth. Nonl. Anal.* 10 (1997), 47–78.
[26] Y. Long, Hyperbolic closed characteristics on compact convex smooth hypersurfaces. *J. Diff. Equa.* 150 (1998), 227–249.
[27] Y. Long, Bott formula of the Maslov-type index theory. *Pacific J. Math.* 187 (1999), 113–149.
[28] Y. Long, Multiple periodic points of the Poincaré map of Lagrangian systems on tori. *Math. Z.* 233 (2000) 443–470.
[29] Y. Long, Precise iteration formulae of the Maslov-type index theory and ellipticity of closed characteristics. *Advances in Math.* 154 (2000), 76–131.

[30] Y. Long, Index Theory for Symplectic Paths with Applications. Progress in Math. 207, Birkhäuser. Basel. 2002.
[31] Y. Long & T. An, Indexing the domains of instability for Hamiltonian systems. *NoDEA*. 5 (1998) 461–478.
[32] Y. Long & D. Dong, Normal forms of symplectic matrices. *Acta Math.Sinica*. 16 (2000) 237–260.
[33] Y. Long & E. Zehnder, Morse theory for forced oscillations of asymptotically linear Hamiltonian systems. In *Stoc. Proc. Phys. and Geom.* S. Albeverio et al. ed. World Sci. (1990) 528–563.
[34] Y. Long & C. Zhu, Closed characteristics on compact convex hypersurfaces in $\mathbf{R}^{2n}$. *Annals of Math.* 155 (2002) 317–368.
[35] P. H. Rabinowitz, Periodic solutions of Hamiltonian systems. *Comm. Pure Appl. Math.* 31 (1978) 157–184.
[36] P. H. Rabinowitz, Minimax methods in critical point theory with applications to differential equations. *CBMS Regional Conf. Ser. in Math.* no.65. Amer. Math. Soc. 1986.
[37] D. Salamon & E. Zehnder, Morse theory for periodic solutions of Hamiltonian systems and the Maslov index. *Comm. Pure and Appl. Math.* 45. (1992). 1303–1360.
[38] H. Seifert, Periodischer Bewegungen mechanischen systeme. *Math. Z.* 51 (1948), 197–216.
[39] C. Viterbo, A new obstruction to embedding Lagrangian tori. *Invent. Math.* 100 (1990), 301–320.
[40] A. Weinstein, Periodic orbits for convex Hamiltonian systems. *Ann. of Math.* 108. (1978). 507–518.

ICM 2002 · Vol. II · 315–321

Some Applications of Collapsing with Bounded Curvature

Anton Petrunin*

Abstract

In my talk I will discuss the following results which were obtained in joint work with Wilderich Tuschmann.

1. For any given numbers m, C and D, the class of m-dimensional simply connected closed smooth manifolds with finite second homotopy groups which admit a Riemannian metric with sectional curvature $|K| \leq C$ and diameter $\leq D$ contains only finitely many diffeomorphism types.

2. Given any m and any $\delta > 0$, there exists a positive constant $i_0 = i_0(m, \delta) > 0$ such that the injectivity radius of any simply connected compact m-dimensional Riemannian manifold with finite second homotopy group and Ricci curvature $Ric \geq \delta$, $K \leq 1$, is bounded from below by $i_0(m, \delta)$.

I also intend to discuss Riemannian megafolds, a generalized notion of Riemannian manifolds, and their use and usefulness in the proof of these results.

2000 Mathematics Subject Classification: 53C.

This note is about a couple of applications and variations of techniques developed in [CFG], which we found jointly with W. Tuschmann. Namely I will talk about injectivity radius estimates for positive pinching, a generalized notion of manifolds, and finiteness theorems for Riemannian manifolds with bounded curvature. The purpose of this note is to give an informal explanation of ideas in these proofs and for more details I refer the reader to [PT].

1. Injectivity radius estimates and megafolds

Is it true that positive pinching of the sectional curvatures of a simply connected manifold implies some lower positive bound for the injectivity radius, which does not depend on the manifold? For dimension = 3 this was proved by Burago and Toponogov [BT]. More generally, they proved the following:

*Department of Mathematics, PSU, University Park, PA 16802, USA. E-mail: petruninth.psu.edu

Theorem A. *Given any $\delta > 0$, there exists a positive constant $i_0 = i_0(\delta) > 0$ such that the injectivity radius of any simply connected compact 3-dimensional Riemannian manifold with Ricc$\geq \delta$, $K \leq 1$, is bounded from below by i_0.*

Moreover they made a conjecture that this result should be also true for higher dimensions. Later on some new examples of manifolds with positively pinched curvature were found by Alloff and Wallach, Eschenburg and Bazaikin ([AW], [E], [B]) which disprove this conjecture in general, but since then closely related conjectures appeared on almost each list of open problems in Riemannian geometry. The theorem which we proved can be formulated as follows:

Theorem B. *Given any m and any $\delta > 0$, there exists a positive constant $i_0 = i_0(m, \delta) > 0$ such that the injectivity radius of any simply connected compact m-dimensional Riemannian manifold with finite second homotopy group and Ricc$\geq \delta$, $K \leq 1$, is bounded from below by $i_0(m, \delta)$.*

Theorem B generalizes the Burago-Toponogov Theorem A to arbitrary dimensions and is also in even dimensions interesting, since there is no Synge theorem for positive Ricci curvature. For sectional curvature pinching a similar result was obtained independently by Fang and Rong [FR].

Now I will turn to one proof of this statement which is described in the appendix of [PT] (The main part of paper contains an other proof). This proof makes use of a generalized notion of Riemannian manifold, which was also described by Gromov in the end of section 8_+ of [G3], and employs a "tangential" version of Gromov-Hausdorff convergence. Here I will just give an informal analogy which describes this notion. The formal aspects and all further details can be found in [PT].

One may think about a manifold as a set of charts and glueing mappings. For a Riemannian manifold, denoting the disjoint union of all charts with the pulled back metrics by (U, g), the set of all glueing maps defines an isometric pseudo-group action by a pseudogroup G on (U, g). Here is the definition of a pseudogroup action:

Definition. *A pseudogroup action (or pseudogroup of transformations) on a manifold M is given by a set G of pairs of the form $p = (D_p, \bar{p})$, where D_p is an open subset of M and $\bar{p}$ is a homeomorphism $D_p \to M$, so that the following properties hold:*

(1) $p, q \in G$ implies $p \circ q = (\bar{q}^{-1}(D_p \cap \bar{q}(D_p)), \bar{p} \circ \bar{q}) \in G$;

(2) $p \in G$ implies $p^{-1} = (\bar{p}(D_p), \bar{p}^{-1}) \in G$;

(3) $(M, id) \in G$;

(4) if $\bar{p}$ is a homeomorphism from an open set $D \subset M$ into M and $D = \bigcup_\alpha D_\alpha$, where D_α are open sets in M, then the property $(D, \bar{p}) \in G$ is equivalent to $(D_\alpha, \bar{p}|_{D_\alpha}) \in G$ for any α.

We call the pseudo-group action natural if in addition the following is true:

(i)' If $(D, \bar{p}) \in G$ and $\bar{p}$ can be extended as a continuous map to a boundary point $x \in \partial D$, then there is an element $(D', \bar{p}') \in G$ such that $x \in D'$, $D \subset D'$ and $\bar{p}'|_D = \bar{p}$.

To form a manifold this action must be in addition properly discontinuous

and free. If it just properly discontinuous then we obtain an orbifold. In the case of a general (isometric!) pseudogroup action we obtain a *it Riemannian megafold* (cf. [PT]). The megafold which is obtained this way will be denoted by $(\mathcal{M}, g) = ((U, g) : G)$.

Now we come to the main notion of this section:

Definition. *A sequence of Riemannian megafolds $(\mathcal{M}_n, g_n)$ is said to Grothendieck-Lipschitz converge (GL-converge) to a Riemannian megafold $(\mathcal{M}, g)$ if there are representations $(\mathcal{M}_n, g_n) = ((U_n, g_n) : G_n)$ and $(\mathcal{M}, g) = ((U, g) : G)$ such that*

(a) The (U_n, g_n) Lipschitz converge to (U, g), and

(b) For some sequence $\epsilon_n \to 0$ there is a sequence of $e^{\pm\epsilon_n}$-bi-Lipschitz homeomorphisms $h_n : (U_n, g_n) \to (U, g)$ such that the pseudogroup actions on $\{(U_n, g_n)\}$ converge (with respect to the homeomorphisms h_n) to a pseudogroup action on $\{(U, g)\}$.

I.e., for any converging sequence of elements $p_{n_k} \in G_{n_k}(U_{n_k}, \mathcal{M}_{n_k})$ there exists a sequence $p_n \in G_n$ which converges to the same local isometry on U, and the pseudogroup of all such limits, acting on U, coincides with the pseudogroup action $G(U, \mathcal{M})$.

Here are two simple examples of GL-convergence:

Consider the sequence of Riemannian manifolds $S^1_\epsilon \times \mathbb{R}$, which for $\epsilon \to 0$ Gromov-Hausdorff converge to $\mathbb{R}$. Then this sequence converges in the GL-topology to a Riemannian megafold $\mathcal{M}$, which can described as follows: It is covered by one single chart $U = \mathbb{R}^2$, and the pseudogroup G simply consists of all vertical shifts of $\mathbb{R}^2$. I.e., $\mathcal{M}$ is nothing but $(\mathbb{R}^2 : \mathbb{R})$ where $\mathbb{R}$ acts by parallel translations. (Note that $(\mathbb{R}^2 : \mathbb{R}) \neq \mathbb{R}^2/\mathbb{R}$, these megafolds even have different dimensions!)

The Berger spheres, as they Gromov-Hausdorff collapse to S^2, converge in Grothendieck-Lipschitz topology to the Riemannian megafold $(S^2 \times \mathbb{R} : \mathbb{R})$. Here $\mathbb{R}$ acts by parallel shifts of $S^2 \times \mathbb{R}$.

Notice that a Riemannian metric on a megafold $((U, g) : G)$ defines a pseudometric on the set of G orbits. In particular one has that the diameter of a Riemannian megafold is well defined. Now here is the basic result, whose proof is obvious from the definitions:

Theorem C. *The set of Riemannian m-manifolds (megafolds) with bounded sectional curvature $|K| \leq 1$ and diameter $\leq D$ is precompact (compact) in the Grothendieck-Lipschitz topology.*

Now let us state some natural questions which arise from this theorem:

1. Which Riemannian megafolds can be approximated by manifolds with bounded curvature and diameter?

Note that the infinitesimal motions of the pseudogroup G give rise to a Lie algebra of Killing fields on a megafold (U, g) from which one can recover an isometric local action of a connected Lie group on (U, g). Let us call this group G_o. Then G_o is obviously an invariant of the megafold, i.e., does not depend on a particular representation $(U : G)$. It follows now from [CFG] that if $(\mathcal{M}, g)$ is a limit of

Riemannian manifolds with bounded curvature, then $G_o(\mathcal{M})$ must be nilpotent. A direct construction moreover shows that this condition is also sufficent.

(Note that since a pure N-structure on a simply connected manifold is given by a torus action, one also has the following: If a megafold can be approximated by simply connected manifolds with bounded curvature, then $G_o(\mathcal{M}) = \mathbb{R}^k$.)

2. How can one recover the Gromov-Hausdorff limit space from a Grothendieck-Lipschitz limit?

Let $\mathcal{M} = ((U, g) : G)$ be a GL-limit of Riemannian manifolds. The GH-limit is the space of G orbits with the induced metric, in other words: The Gromov-Hausdorff limit is nothing but $(U, g)/G$.

Riemannian megafolds are actually not that general objects as they might seem at first sight. Indeed, given a Riemannian megafold $(\mathcal{M}, g)$ we can consider its orthonormal frame bundle $(F\mathcal{M}, \tilde{g})$, equipped with the induced metric. Now consider some representation of it, say, $(F\mathcal{M}, \tilde{g}) = ((U, \tilde{g}) : G)$. Then the G pseudogroup action is free on U, so that its closure $\bar{G}$ also acts freely. Therefore the corresponding factor, equipped with the induced metric, is a Riemannian manifold $Y = (U/\bar{G}, \bar{g})$, and there is a Riemannian submersion $(F\mathcal{M}, \tilde{g}) \to (U/\bar{G}, \bar{g})$ whose fibre is G_o/Γ_o, where Γ_o is a dense subgroup of G_o (Roughly speaking, Γ_o is generated by the intersections of G_o and G). If we assume that $\mathcal{M}$ is simply connected, then $G_o = \mathbb{R}^k$ and Γ_o is the homotopy sequence image of $\pi_2(Y)$. In particular, the dimension of the free part of $\pi_2(Y)$ is at least $k + 1$.

Notice that for Riemannian megafolds one can define the de Rham complex just as well as for manifolds. (In fact I am not aware of a single notion or theorem in Riemannian geometry which does not admit a straightforward generalization to Riemannian megafolds!) From the above characterization of Riemannian megafolds it is not hard to obtain the following:

Theorem D. *Let M_n be a sequence of compact simply connected Riemannian m-manifolds with bounded curvatures and diameters and $H^2_{dR}(M_n) = 0$ which Grothendieck-Lipschitz converges to a Riemannian megafold $(\mathcal{M}, g)$.*

Then $\mathcal{M}$ is either a Riemannian manifold and the manifolds M_n converge to $\mathcal{M}$ in the Lipschitz sense, or $H^1_{dR}(\mathcal{M}) \neq 0$.

It is in particular straightforward to show that if $\mathrm{Ricc}(\mathcal{M}) > 0$, then $H^1_{dR}(\mathcal{M}) = 0$. Moreover, a Grothendieck-Lipschitz limit of manifolds with uniformly bounded sectional curvatures and $\mathrm{Ricc} \geq \delta > 0$ is a Riemannian megafold with $\mathrm{Ricc} \geq \delta > 0$.

Now we can prove Theorem B: Assume it is wrong. Then we can find a collapsing sequence of simply connected manifolds with finite π_2 and positive Ricci-pinching, and we obtain a megafold with $H^1_{dR} \neq 0$ as a GL-limit. Applying the Bochner formula for 1-forms on this megafold, we obtain a contradiction.

2. Finiteness theorems

The following result appeared as a co-product of the theorem above, and it came as a nice surprise. Let me first formulate this finiteness results from [PT]:

Theorem E (The π_2-Finiteness Theorem). *For given m, C and D, there is only a finite number of diffeomorphism types of simply connected closed m-dimensional manifolds M with finite second homotopy groups which admit Riemannian metrics with sectional curvature $|K(M)| \leq C$ and diameter $diam(M) \leq D$.*

Theorem F (A "classification" of simply connected closed manifolds). *For given m, C and D, there exists a finite number of closed smooth simply connected manifolds E_i with finite second homotopy groups such that any simply connected closed m-dimensional manifold M admitting a Riemannian metric with sectional curvature $|K(M)| \leq C$ and diameter $diam(M) \leq D$ is diffeomorphic to a factor space $M = E_i/T^{k_i}$, where $0 \leq k_i = b_2(M) = dim\ E_i - m$ and T^{k_i} acts freely on E_i.*

Here is a short account of other finiteness results which only require volume, curvature, and diameter bounds: For manifolds M of a given fixed dimension m, the conditions

- $vol(M) \geq v > 0$, $|K(M)| \leq C$ and $\mathrm{diam}(M) \leq D$ imply finiteness of diffeomorphism types (Cheeger ([C]) 1970); this conclusion continues to hold for $vol(M) \geq v > 0$, $\int_M |R|^{m/2} \leq C$, $|\mathrm{Ric}_M| \leq C'$, $\mathrm{diam}(M) \leq D$ (Anderson and Cheeger ([AC1]) 1991);
- $vol(M) \geq v > 0$, $K(M) \geq C$ $\mathrm{diam}(M) \leq D$ imply finiteness of homeomorphism types (Grove-Petersen ([GP]) 1988, Grove-Petersen-Wu ([GPW]) 1990); Perelman, ([Pe]) 1992) (if in addition $m > 4$, these conditions imply finiteness of diffeomorphism types) and Lipschitz homeomorphism types (Perelman, unpublished);
- $K(M) \geq C$ and $\mathrm{diam}(M) \leq D$ imply a uniform bound for the total Betti number (Gromov [G1] 1981).

The π_2-Finiteness Theorem requires two-sided bounds on curvature, but no lower uniform volume bound. Thus, in spirit it is somewhere between Cheeger's Finiteness and Gromov's Betti number Theorem.

Each of the above results has (at least) two quite different proofs, the original one and one which uses Alexandrov techniques. (For Gromov's Betti number theorem we made such a proof recently, jointly with V. Kapovich and it turned out that one can even give an upper estimate for the total number of critical points of a Morse function on such a manifold, which due to the Morse inequality is a stronger condition.) Let me now explain roughly this second way of proving of such theorems:

I will take Cheeger's theorem as an example: Assume it is wrong. Then there is an infinite number of non-diffeomorphic manifolds with bounded curvature, diameter and a lower bound on the volume. Then due to Gromov's compactness theorem a subsequence of them has a limit. Then, due to the volume bound, this limit space has the same dimension, and is in fact just little worse than Riemannian; it is a manifold with a smooth structure and curvature bounded in the sense of Alexandrov. Then one only has to prove the stability result, i.e. one has to prove that starting from some big number all manifolds are diffeomorphic to the limit

space. In the case of two-sided curvature bound it is really simple, and for just lower curvature bound it is already a hard theorem of Perelman, but still it works along this lines.

Now for both of these proofs it is very important to have a uniform lower positive volume bound to prevent collapsing. In fact, if one removes this bound then it is not hard to construct infinite sequence of non-diffeomorphic manifolds. This holds for two-sided bounded as well as for lower curvature bound. And if we would try to prove it the same way as before we would get a limit space of possibly smaller dimension. Therefore the stability result can not hold this way.

This partly explains why Theorem E looks a bit surprising, we add one topological condition and get real finiteness result. The proof can go along the same lines. Take a sequence of nondiffeomorphic Riemannian manifold (M_n, g_n), by Gromov's compactness theorem we have a limit space (for some subsequence) X. The sequence must collapse, otherwise the same arguments as before would work. Since the M_n are simply connected, from [CFG] we have that collapsing takes place along some T^k-orbits of some T^k-action.

Now assume for simplicity that X is a manifold and $\pi_2(M_n) = 0$. Then all M_n are diffeomorphic to T^k bundles over X. Since the M_n are simply connected so is X. Therefore the diffeomorphism type of M_n depends only on the Euler class e_n which in this case can be interpreted as the following mapping:

$$0 = \pi_2(M_n) \to \pi_2(X) \xrightarrow{e_n} \pi_1(T^k) \to \pi_1(M_n) = 0.$$

Therefore e_n isan isomorphism between two groups and up to automorphisms of T^k all possible Euler classes e_n are the same. In particular, for large n all M_n are diffeomorphic.

That is not quite a proof since we had made quite strong assumptions on the way. But it turns out that the general case can be ruled out using a few already standard tricks from [CFG] and [GK], namely, by passing to the frame bundles FM_n and by conjugating group actions.

References

[AW] S. Aloff; N. R. Wallach, *An infinite family of 7-manifolds admitting positively curved Riemannian structures*, Bull. Amer. Math. Soc. **81** (1975), 93–97.

[B] Ya. V. Basaikin, *On a certain family of closed 13-dimensional manifolds of positive curvature*, Siberian Mathematical Journal **37:6** (1996).

[BT] Y. Burago; V. A. Toponogov, *On three-dimensional Riemannian spaces with curvature bounded above*, Matematicheskie Zametki **13** (1973), 881–887.

[C] J. Cheeger, *Finiteness theorems for Riemannian manifolds*, Amer. J. Math. **92** (1970), 61–74.

[CFG] J. Cheeger; K. Fukaya; M. Gromov, *Nilpotent structures and invariant metrics on collapsed manifolds*, J. A.M.S **5** (1992), 327–372.

[Es] J.-H. Eschenburg, *New examples of manifolds with strictly positive curvature*, Invent. math. **66** (1982), 469–480.

[FR] F. Fang; X. Rong, *Positive Pinching, volume and second Betti number*, GAFA (Geometric and functional analysis **9** (1999).

[G1] M. Gromov, *Curvature, diameter and Betti numbers*, Comment. Math. Helv. **56** (1981), 179–195.

[G2] M. Gromov, *Stability and Pinching*, Seminare di Geometria, Giornate di Topologia e geometria delle varieta. Universita degli Studi di Bologna (1992), 55–97.

[G3] M. Gromov, *"Metric Structures for Riemannian and Non-Riemannian spaces"*, Birkhäuser, Basel, (1999).

[GK] K. Grove; H. Karcher, *How to conjugate C^1-close group actions*, Math. Z. **132** (1973), 11–20.

[GPW] K. Grove; P. Petersen; J. Wu, *Controlled topology in geometry*, Invent. Math. **99** (1990), 205–213; Erratum: Invent. Math. 104 (1991), 221–222.

[KS1] W. Klingenberg; T. Sakai, *Injectivity radius estimates for 1/4-pinched manifolds*, Arch. Math. **34** (1980), 371–376.

[PRT] A. Petrunin; X. Rong; W. Tuschmann, *Collapsing vs. Positive Pinching*, GAFA (Geometric and functional analysis) **9** (1999), 699–735.

[PT] A. Petrunin and W. Tuschmann, *Diffeomorphism finiteness, positive pinching, and second homotopy*, GAFA (Geometric and functional analysis) **9** (1999), 736–774.

[PT2] A. Petrunin and W. Tuschmann, *Asymptotical flatness and cone structure at infinity*, Math. Ann. **321** (2001), 775–788.

ICM 2002 · Vol. II · 323–338

Collapsed Riemannian Manifolds with Bounded Sectional Curvature*

Xiaochun Rong†

Abstract

One of the most important developments in Riemannian geometry over the last two decades is the structure theory of Cheeger-Fukaya-Gromov, for manifolds M^n of bounded sectional curvature, say $|K_{M^n}| \leq 1$, which are sufficiently collapsed. Roughly, M^n is called ϵ-collapsed, if it appears to have dimension less than n, unless the metric is rescaled by a factor $\geq \epsilon^{-1}$. For example, a very thin cylinder is very collapsed (although its curvature vanishes identically).

If one fixes ϵ and in addition, a bound, d, on the diameter, then in each dimension, there are only finitely many manifolds, which are not ϵ-collapsed. The basic result of collapsing theory states the existence of a constant, $\epsilon(n) > 0$, such that a manifold which is ϵ-collapsed, for $\epsilon \leq \epsilon(n)$, has a particular kind of singular fibration structure with flat (or "almost flat") fibers. The fibers lie in the ϵ-collapsed directions.

The first nontrivial collapsing with bounded curvature, arose in a sequence of metrics on the 3-sphere constructed by M. Berger. The first major result on the collapsed manifolds (still a corner stone of the theory) is M. Gromov's description of "almost flat manifolds" i.e. manifolds admitting a sequence of metrics with curvature and diameter going to zero. Gromov showed that such manifolds are infranilmanifolds.

We will survey the main development of the collapsing theory and its applications to Riemannian geometry since 1990. The common starting point is the existence of the above mentioned singular fibration structure. Many new geometrical and topological constraints of collapsed metrics have been discovered that are accompanied with new ideas and techniques as well as tools from related fields, and light has been shed on some classical problems and conjectures, which do not, on the face of it, involve collapsing. Substantial progress has been made on manifolds with non-positive curvature, on positively pinched manifolds, collapsed manifolds with an a priori diameter bound, and subclasses whose members satisfy additional topological conditions e.g. 2-connectedness.

2000 Mathematics Subject Classification: 53C.

*Supported partially by NSF Grant DMS 0203164 and a research found from Beijing Normal University.

†Rutgers University, New Brunswick, NJ 08903, USA and Beijing Normal University, Beijing 100875, China. E-mail: rong@math.rutgers.edu

One of the most important developments in Riemannian geometry over the last two decades is the structure theory of Cheeger-Fukaya-Gromov for manifolds M^n of bounded sectional curvature, say $|\sec_{M^n}| \leq 1$, which are sufficiently collapsed. Roughly, M^n is called ϵ-collapsed, if it appears to have dimension less than n, unless the metric is rescaled by a factor $\geq \epsilon^{-1}$.

For scaling reasons, collapsing and boundedness of tend to oppose one another. Nevertheless, very collapsed manifolds with bounded curvature do in fact exist. For example, a very thin cylinder is very collapsed, although its curvature vanishes identically.

If one fixes ϵ and in addition, a bound, d, on the diameter, then in each dimension, there only finitely many manifolds, which are not ϵ-collapsed; see [Ch]. The basic result of collapsing theory states the existence of a constant $\epsilon(n) > 0$, such that a manifold which is ϵ-collapsed, for $\epsilon \leq \epsilon(n)$, has a particular kind of singular fibration structure with flat (or "almost flat") fibers. The fibers lie in the ϵ-collapsed directions; see [CG1,2], [CFG], [Fu1-3].

The first nontrivial example of a collapsing sequence with bounded curvature (described in more detail below) was constructed by M. Berger in 1962; see [CFG]. The first major result on the collapsed manifolds (still a cornerstone of the theory) was M. Gromov's characterization of "almost flat manifolds" i.e. manifolds admitting a sequence of metrics with curvature and diameter going to zero. Gromov showed that such manifolds are infranil. Later in [Ru], they were shown to actually be nilmanifolds; compare [GMR].

We will survey the development of collapsing theory and its applications to Riemannian geometry since 1990; compare [Fu4]. The common starting point for all of these is the above mentioned singular fibration structure. However, new techniques have been introduced and tools from related fields have been brought in. As a consequence, light has been shed on some classical problems and conjectures whose statements do not involve collapsing. Specifically, substantial progress has been made on manifolds with nonpositive curvature, on positively pinched manifolds, collapsed manifolds with an a priori diameter bound, and subclasses of manifolds whose members satisfy additional topological conditions e.g. 2-connectedness.

1. Collapsed manifolds of bounded sectional curvature

Convention: unless otherwise specified, "collapsing" refers to a sequence of Riemannian manifolds with sectional curvature bounded in absolute value by one and injectivity radii uniformly converge to zero, while "convergence" means "convergence with respect to the Gromov-Hausdorff distance.

Recall that a map from a metric space (X, d_X) to a metric space (Y, d_Y) is called an ϵ-Gromov-Hausdorff approximation, if of $f(X)$ is ϵ-dense in Y and if $|d_X(x, x') - d_Y(f(x), f(x'))| < \epsilon$. The Gromov-Hausdorff distance between two (compact) metric spaces is the infimum of ϵ as above, for all possible ϵ-Gromov-Hausdorff approximations from X to Y and vice versa. (To be more precise, one should say "pseudo-distance", since isometric metric spaces have distance zero.)

The collection of all compact metric spaces is complete with respect to the Gromov-Hausdorff distance.

a. Flat manifolds, collapsing by scaling and torus actions

For fixed (M, g), the family, $\{(M, \epsilon^2 g)\}$ converges to a point as $\epsilon \to 0$. However, if the curvature is not identically zero, it blows up. On the other hand, for any compact flat manifold, (M, g), the the manifolds, $(M, \epsilon^2 g)$ continue to be flat. More generally, if (M, g) is a (possibly nonflat) manifold with an isometric torus T^k-action for which all T^k-orbits have the same dimension, then one obtains a collapsing sequence by rescaling g along the orbits i.e. by putting $g_\epsilon = \epsilon^2 g_0 \oplus g_0^\perp$, where g_0 is the restriction of g to the tangent space of a T^k-orbit and $g_0^\perp$ is the orthogonal complement. A computation shows that g_ϵ has bounded sectional curvature independent of ϵ. The collapse constructed by Berger in 1962 was of this type. In his example, M^3 is the unit 3-sphere and the S^1 action is by rotation in the fibers of the Hopf fibration $S^1 \to S^3 \to S^2$. The limit of this collapse is the 2-sphere with a metric of constant curvature $\equiv 4$; see [Pet].

More generally, a collapsing construction has been given by Cheeger-Gromov for manifolds which admit certain mutually compatible local torus actions (possibly by tori of different dimensions) for which all orbits have positive dimension; see the notion of *F-structure* given below and (1.2.1). As above, for each individual local torus action, one obtains locally defined collapsing sequence. The problem is to patch together these local collapsings. If the orbits are not all of the same dimension, the patching requires a suitable scaling of the metric (by a large constant) in the transition regions between orbits of different dimensions; see [CG1]. Hence, in contrast to the Berger example, in general the diameters of such nontrivially patched collapsings necessarily go to infinity.

b. Almost flat manifolds and collapsing by inhomogeneous scaling

Although a compact nilmanifold (based on a nonablian nilpotent Lie group) admits no flat metric, a sequence metrics on such a manifold which collapses to a point can be constructed by a suitable inhomogeneous scaling process; see [Gr1]. As an example, regard a compact nilmanifold M^3 as the total space of a principle circle bundle over a torus. A canonical metric g on M^3 splits into horizontal and vertical complements, $g = g_h \oplus g_h^\perp$. Then $g_\epsilon = (\epsilon g_h) \oplus (\epsilon^2 g_h^\perp)$ has bounded sectional curvature independent of ϵ, while (M^3, g_ϵ) converges to a point. The inhomogeneity of the scaling is essential in order for the curvature to remain bounded; compare Theorem 3.4.

c. Positive rank F-structure and collapsed manifolds

The notion of an F-structure may be viewed as a generalization of that of a torus action. An F-structure $\mathcal{F}$ on a manifold is defined by an atlas $\mathcal{F} = \{(V_i, U_i, T^{k_i})\}$, satisfying the following conditions:

(1.1.1) $\{U_i\}$ is a locally finite open cover for M.

(1.1.2) $\pi_i : V_i \to U_i$ is a finite normal covering and V_i admits an effective torus T^{k_i}-action such that it extends to a $\pi_1(U_i) \ltimes T^{k_i}$-action.

(1.1.3) If $U_i \cap U_j \neq \emptyset$, then $\pi_i^{-1}(U_i \cap U_j)$ and $\pi_j^{-1}(U_i \cap U_j)$ have a common finite covering on which the lifting $T^{k_i}-$ and T^{k_j}-actions commute.

If $k_i = k$, for all i, then $\mathcal{F}$ is called *pure*. Otherwise, $\mathcal{F}$ is called *mixed*. The compatibility condition, (1.1.3), implies that M decomposes into *orbits*. (an orbit at a point is the smallest set containing all the projections of the T^{k_i}-orbits at the point.) The minimal dimension of all such orbits is called the *rank of* $\mathcal{F}$. An orbit is called *regular*, if it has a tubular neighborhood in which the orbits form a fibration. Otherwise, it is called *singular*. An F-structure $\mathcal{F}$ is called *polarized* if all T^{k_i}-actions are almost free. An F-structure is called *injective* (resp. semi-injective) if the inclusion of any orbit to M induces an injective (resp. nontrivial) map on the fundamental groups.

A *Cr-structure* is an injective F-structure with an atlas that satisfies two additional properties: i) $V_i = D_i \times T^{k_i}$ and T^{k_i} acts on V_i by the multiplication. ii) If $U_i \cap U_j \neq \emptyset$, then $k_i < k_j$ or vice versa; see [Bu1]. This notion arises in the context of nonpositive curvature.

A metric is called an *$\mathcal{F}$-invariant* (or simply invariant), if the local T^{k_i}-actions are isomeric. For any F-structure, there exists an invariant metric.

A manifold may not admit any nontrivial F-structure; compare Corollary 2.5. In fact, a simple necessary condition for a closed manifold M^{2n} to admit a positive rank F-structure is the vanishing of its Euler characteristic; see [CG1].

A necessary and sufficient condition for the existence of a collapsing sequence of metrics is the existence of an F-structure of positive rank; see [CG1], [CG2].

Theorem 1.2 (Collapsing and F-structure of positive rank). ([CG1,2]) *Let M be a manifold without boundary.*
(1.2.1) If M admits a positive rank (resp. polarized) F-structure, then M admits a continuous one-parameter family of invariant metrics g_ϵ such that $|sec_{g_\epsilon}| \leq 1$ and the injectivity radius (resp. volume) of g_ϵ converges uniformly to zero as $\epsilon \to 0$.
(1.2.2) There exists a constant $\epsilon(n)$ (the critical injectivity radius) such that if M^n admits a metric g with $|sec_g| \leq 1$ and the injectivity radius is less than $\epsilon(n)$ everywhere, then M admits a positive rank F-structure almost compatible with the metric.

The F-structure in (1.2.2) is actually a substructure of a so called *nilpotent Killing structure on M* whose orbits are infra-nilmanifolds; see [CFG] and compare to Theorem 3.5. Such an infra- nilmanifold orbit at a point contains *all* sufficiently collapsed directions of the metric; the orbit of its sub F-structure, which is defined by the 'center' of the infra-nilmanifold, only contains the most collapsed directions comparable to the injectivity radius at a point. A unsolved problem pertaining to nilpotent structures is whether a collapse as in (1.2.1) can be constructed for which the diameters of the nil-orbits converge uniformly to zero (as holds for F-structures).

The construction of the F-structure in (1.2.2) relies only on the local geometry. Hence, (1.2.2) can be applied to a collapsed region in a complete manifold of bounded sectional curvature. In this way, for such a manifold, one obtains a *thick-thin* decomposition, in which the thin part carries an F-structure of positive rank; see [CFG].

Theorem 1.2 has been the starting point for many subsequent investigations of collapsing in various situations. The guiding principle is that additional geometrical properties of a collapsing should be mirrored in properties of its associated F-structure, which in turn, puts constraints on the topology. For instance, if a collapsing satisfies additional geometrical conditions such as: i) volume small, ii) uniformly bounded diameter, iii) nonpositive curvature, iv) positive pinched curvature, v) *bounded covering geometry* i.e. the injectivity radii of the Riemannian universal covering has a uniform positive lower bound, then one may expect corresponding topological properties of the F-structure such as: i) existence of a polarization, ii) pureness, iii) existence of a Cr-structure, iv) the existence of a circle orbit, v) injective F-structure. Results on such correspondences and their applications will occupy the rest of this paper.

d. Topological invariants associated to a volume collapse

The existence of a sufficiently (injectivity radius) collapsed metric as in (1.2.2) imposes constraints on the underlying topology. For instance, the *simplicial volume* of M vanishes; see [Gr3]. As mentioned earlier, for a closed M^{2n}, the Euler characteristic of M^{2n} also vanishes; see [CFG].

In this subsection, we focus on some topological invariants associated to certain (partially) volume collapsed metrics: the *minimal volume*, the *L^2-signature* and the *limiting η-invariant*; see below.

The minimal volume, $\mathrm{MinVol}(M)$, of M, is the infimum of the volumes over all complete metrics with $|\sec_M| \leq 1$. Clearly, $\mathrm{MinVol}(M)$ is a topological invariant. Gromov conjectured that there exists a constant $\epsilon(n) > 0$ such that $\mathrm{MinVol}(M^n) < \epsilon(n)$ implies that $\mathrm{MinVol}(M^n) = 0$ (the gap conjecture for minimal volume). By Theorem 1.2, it would suffice to show that a sufficiently volume collapsed manifold admits a polarized F-structure. On a 3-manifold, any positive rank F-structure has a polarized substructure and thus Theorem 1.2 implies Gromov's gap conjecture in dimension 3. However, for $n \geq 4$, there are n-manifolds which admit a positive rank F-structure but which admit no polarized F-structure; see [CG1].

Theorem 1.3 (Volume collapse and Polarized F-structure). ([Ro2]) *There is a constant $\epsilon > 0$ such that if $MinVol(M^4) < \epsilon$, then M^4 admits a polarized F-structure and thus $MinVol(M^4) = 0$.*

For a complete open manifold with bounded sectional curvature and finite volume (necessarily volume collapsed near infinity), the integral of an invariant polynomial of the curvature form may depend on the particular metric; see [CG3]. It is of interest to find a class of metrics for which integral of characteristic forms have a topological interpretation. Cheeger-Gromov showed that for any open complete manifold M^{4k} of finite volume and bounded covering geometry outside some compact subset, the integral of the Hirzebruch signature form over M^{4k} is independent of the metric; see [CG3] and the references therein. Cheeger-Gromov showed that this integral is equal to the so called L_2-signature and conjectured that it can take only rational values. (The notion of L_2-signature, whose definition involves the concept of Von Neumann dimension, was first introduced by Atiyah and Singer in the context of coverings of compact manifolds.)

Theorem 1.4 (Rationality of geometric signature). ([Ro3]) *If an open complete manifold, M^4, of finite volume has bounded covering geometry outside a compact subset, then the integral of the Hirzebruch signature form over M^4 is a rational number.*

The main idea is to show that M^4 admits a polarized F-structure $\mathcal{F}$ outside some compact subset and an exhaustion by compact submanifolds, M_i^4, such that the restriction of $\mathcal{F}$ to the boundary of M_i^n is injective. The integral over M^4 is the limit of the integrals over M_i^4, to which we apply the Atiyah-Patodi-Singer formula to reduce to showing the rationality of the limit of the η-invariant terms. By making use of the special property of $\mathcal{F}$ and Theorem 1.5 below, we are able to conclude that the limit of the η-invariant term is rational.

Cheeger-Gromov showed that if a sequence of volume collapsed metrics on a closed manifold N^{4n-1} have bounded covering geometry, then the sequence of the associated η-invariants converges and the limit is independent of the particular sequence of such metrics. They conjectured that the limit is rational.

Theorem 1.5 (Rationality of limiting η-invariants). ([Ro1]) *If a closed manifold N^3 admits a sequence of volume collapsed metrics with bounded covering geometry, then N^3 admits an injective F-structure and the limit of the η-invariants is rational.*

The idea is to show that N^3 admits an injective F-structure $\mathcal{F}$. For an injective F-structure, the collapsing constructed in (1.2.1) has bounded covering geometry and may be used to compute the limit. Results from 3-manifold topology play a role in the proof of the existence of the injective F-structure.

2. Collapsed manifolds with nonpositive sectional curvature

A classical result of Preismann says that for a closed manifold M^n with negative sectional curvature, any abelian subgroup of the fundamental group is cyclic. By bringing in the discrete group technique, Margulis showed that if the metric is normalized such that $-1 \leq \sec_{M^n} \leq 0$, then there exists at least one point at which the injectivity radius is bounded below by a constant $\epsilon(n) > 0$.

The study of the subsequent study of collapsed manifolds with $-1 \leq \sec \leq 0$ may be viewed as an attempt to describe the special circumstances under which the conclusions of the Preismann and Margulis theorem can fail, if the hypothesis is weakened to nonpositive curvature; see [Bu1-3], [CCR1,2], [Eb], [GW], [LY], [Sc].

A collapsed metric with nonpositive curvature tends to be rigid in a precise sense; see (2.2.1) and (2.2.2). Namely, there exists a *canonical* Cr-structure whose orbits are flat totally geodesic submanifolds. Of necessity, the construction of this Cr-structure is global. By contrast, the construction of less precise (but more generally existing) F-structure is local; see [CG2].

Let $M^n = \tilde{M}^n/\Gamma$, where $\tilde{M}^n$ denotes the universal covering space of M^n with the pull-back metric. A *local splitting structure* on a Riemannian manifold is a Γ-equivariant assignment to each point (of an open dense subset of $\tilde{M}^n$) a specified

neighborhood and a specified isometric splitting of this neighborhood, with a non-trivial Euclidean factor. Hence, a necessary condition for a local splitting structure is the existence of a plane of zero curvature, at every point of M^n. A local splitting structure is *abelian* if the projection to M^n of every nontrivial Euclidean factor as above is a closed *embedded* flat submanifold, and n addition, if two projected leaves intersect, then one of them is contained in the other.

Theorem 2.1 (Abelian local splitting structure and Cr-structure). ([CCR1]) *Let M^n be a closed manifold of $-1 \leq sec_{M^n} \leq 0$.*
(2.1.1) If the injectivity radius is smaller than $\epsilon(n) > 0$ everywhere, then M^n admits an abelian local splitting structure.
(2.1.2) If M^n admits an abelian local splitting structure, then it admits a compatible Cr-structure, whose orbits are the flat submanifolds (projected leaves) of the abelian local splitting structure. In particular, $MinVol(M^n) = 0$.

Theorem 2.1 was conjectured by Buyalo, who proved the cases $n = 3, 4$; see [Bu1–3], [Sc].

Let $\tilde{x} \in M^n$. Let $\Gamma_\epsilon(\tilde{x}) \neq 1$ denote the subgroup of Γ generated by those γ whose displacement function, $\delta_\gamma(\tilde{x}) = d(x, \gamma(\tilde{x}))$, satisfies $d(x, \gamma(\tilde{x})) < \epsilon$. (In the application, ϵ is small.) If all $\Gamma_\epsilon(\tilde{x})$ are abelian, then the minimal sets, $\{\mathrm{Min}(\Gamma_\epsilon(\tilde{x}))\}$, of the $\Gamma_\epsilon(\tilde{x})$ give the desired abelian local splitting structure in (2.1.1). In general, $\Gamma_\epsilon(\tilde{x})$ is only Bieberbach. Then, a crucial ingredient in (2.1.1) is the existence of a 'canonical' abelian subgroup of $\Gamma_\epsilon(\tilde{x})$ of finite index consisting of those elements which are *stable* in the sense of [BGS]. In spirit, the proof of (2.1.2) is similar to the construction in [CG2], but the techniques used are quite different.

The following are some specific questions pertaining to abelian local splitting structures:
(2.2.1) If some metric g on M of nonpositive sectional curvature has an abelian local splitting structure, does every nonpositively curved metric also have such a structure?

(2.2.2) If M has a Cr-structure, does every any nonpositively curved metric on M have a compatible local splitting structure?

Note that an affirmative answer to (2.2.1) and (2.2.2) would imply a kind of *semirigidity.* It would imply that all nonpositively curved metrics on M are alike in a precise sense.

Theorem 2.3 (F-structure and local splitting structure). ([CCR2]) *Let X^n, M^n be closed manifolds such that X^n admits a nontrivial F-structure. Let $f : X^n \to M^n$ have nonzero degree. Then every metric of nonpositive sectional curvature on M^n has a local splitting structure.*

We conjecture that if an F-structure has positive rank, then the local splitting structure is abelian. This conjecture, whose proof would provide an affirmative answer to (2.2.2), has been verified in dimension 3 and in some additional special cases; see [CCR2].

We conclude this section with two consequences of Theorem 2.3.

Corollary 2.4 (Generalized Margulis Lemma). ([CCR2]) *Let M^n be a closed manifold of nonpositive sectional curvature. If the Ricci curvature is negative at some point, then for every metric with $|sec| \leq 1$, there is a point with injectivity radius $\geq \delta(n) > 0$.*

Another consequence is a geometric obstruction for a nontrivial F-structure.

Corollary 2.5 (Nonexistence of F-structure). ([CCR2]) *If a closed manifold M admits a metric of nonpositive sectional curvature such that the Ricci curvature is negative at some point, then M does not admit a nontrivial F-structure.*

3. Collapsed manifolds with bounded sectional curvature and diameter

In this section, we discuss the class of collapsed manifolds of bounded sectional curvature whose diameters are also bounded. By the Gromov's compactness theorem, any sequence of such collapsed manifolds contains a convergent subsequence; see [GLP]. Hence, without loss of the generality, we only need to consider convergent collapsing sequences.

(3.1) Let $M_i^n \xrightarrow{d_{GH}} X$ denote a sequence of closed manifolds converging to a *compact* metric space X such that $|\sec_{M_i^n}| \leq 1$ and $\dim(X) < n$.

Main Problem 3.2. For i large, investigate relations between geometry and topology of M_i^n and that of X. The following are some specific problems and questions.

(3.2.1) Find topological obstructions for the existence of M_i^n as in (3.1).

(3.2.2) To what extent is the topology of the M_i^n in (3.1) stable when i is sufficiently large?

(3.2.3) Under what additional conditions is it true that $\{M_i^n\}$ as in (3.1) contains a subsequence of constant diffeomorphism type? If all M_i^n are diffeomorphic, then to what extent do the metrics converge?

Note that by the Cheeger-Gromov convergence theorem, the above problems are well understood in the noncollapsed situation $\dim(X) = n$.

Theorem 3.3 (Convergence). ([Ch], [GLP]) *Let $M_i^n \xrightarrow{d_{GH}} X$ be as in (3.1) except $\dim(X) = n$. Then for i large, M_i^n is diffeomorphic to some fixed M^n which is homeomorphic to X and there are diffeomorphisms, $f_i : M^n \to M_i^n$, such that the pulled back metrics, $f_i^*(g_i)$, converge to a metric, g_∞, in the $C^{1,\alpha}$-topology $(0 < \alpha < 1)$.*

Note that as a consequence of Theorem 3.3, topological stability of a sequence as in (3.1) will immediately yield a corresponding finiteness result in terms of the dimension and bounds on curvature and diameter.

e. Structure of collapsed manifolds with bounded diameter

As described in Section 1, any closed nilmanifold M^n admits metrics collapsing to a point.

Theorem 3.4 (Almost flat manifolds). ([Gr1]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). If* X *is a point, then a finite normal covering space of* M_i^n *of order at most* $c(n)$ *is diffeomorphic to a nilmanifold* N^n/Γ_i *(i large), where* N^n *is the simply connected nilpotent group.*

Theorem 3.4 can be promoted to a description of convergent collapsing sequence, of manifolds, M_i^n, as in (3.1). As mentioned following Theorem 1.2, any sufficiently collapsed manifold admits a nilpotent Killing structure; see [CFG]. Here a bound on diameter forces the nilpotent Killing structure to be pure.

For a closed Riemannian manifold M^n, its frame bundle $F(M^n)$ admits a canonical metric determined by the Riemannian connection up to a choice of a bi-invariant metric on $O(n)$. A fibration, $N/\Gamma \to F(M^n) \to Y$, is called $O(n)$-invariant if the $O(n)$-action on $F(M^n)$ preserves both the fiber N/Γ (a nilmanifold) and the structural group. By the $O(n)$-invariance, $O(n)$ also acts on the base space Y. A canonical metric is invariant if its restriction on each N/Γ is left-invariant. A *pure nilpotent Killing structure* on M is an $O(n)$-invariant fibration on $F(M^n)$ for which the canonical metric is also invariant.

Theorem 3.5 (Fibration). ([CFG]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). Then* $F(M_i^n)$ *equipped with canonical metrics contains a convergent subsequence,* $F(M_i^n) \xrightarrow{d_{GH}} Y$, *and* $F(M_i^n)$ *admits an* $O(n)$*-invariant fibration* $N/\Gamma_i \to F(M_i^n) \to Y$ *for which the canonical metric is* ϵ_i*-close in the* C^1 *sense to some invariant metric, where* $\epsilon_i \to 0$.

The following properties are crucial for the study of particular instances of collapsing as in (3.1).

Proposition 3.6. *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1).*
(3.6.1) (Regularity) ([Ro5]) For any $\epsilon > 0$, M_i^n *admits an invariant metric* g_i *such that* $\min(sec_{M_i^n}) - \epsilon \leq sec_{(M_i^n, g_i)} \leq \max(sec_{M_i^n}) + \epsilon$ *for i large.*
(3.6.2) (Equivariance) ([PT], [GK]) The induced $O(n)$*actions on* Y *from the* $O(n)$*-action on* $F(M_i^n)$ *are* C^1*-close and therefore are all* $O(n)$*-equivariant for i large.*

f. Obstructions to collapsing with bounded diameter

Theorem 3.7 (Polarized F-structure and vanishing minimal volume). ([CR2]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). Then the F-substructure associated to the pure nilpotent Killing structure on* M_i^n *contains a (mixed) polarized F-structure. In particular,* $MinVol(M_i^n) = 0$.

Theorem 3.7 may be viewed as a weak version of the Gromov's gap conjecture. Note that the associated F-structure on M_i^n may not be polarized. The existence of a polarized substructure puts constraints on the singularities of the structure.

Theorem 3.8 (Absence of symplectic structure). ([FR3]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). If* $\pi_1(M_i^n)$ *is finite, then* M_i^n *does not support any symplectic structure.*

The proof of Theorem 3.8 includes a nontrivial extension of the well known fact that any S^1-action on a closed simply connected symplectic manifold which preserves the symplectic structure has a nonempty fixed point set.

A geometric obstruction to the existence of a collapsing sequence in (3.1) is provided by:

Theorem 3.9 (Geometric collapsing obstruction). ([Ro7]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). Then* $\limsup(\max_{M_i^n} Ric_{M_i^n}) \geq 0$.

A key ingredient in the proof is a generalization of a theorem of Bochner asserting that a closed manifold of negative Ricci curvature admits no nontrivial invariant pure F-structure (Bochner's original theorem only guarantees the nonexistence of a nontrivial isometric torus action.)

Theorem 3.10 (Pure injective F-structure). ([CR1]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). If* M_i^n *has bounded covering geometry and* $\pi_1(M_i^n)$ *is torsion free, then for* i *large* M_i^n *admits a pure injective F-structure.*

g. The topological and geometric stability

In this subsection, we address Problems (3.2.2) and (3.2.3). Observe that by the Gromov's Betti number estimate, [Gr2], the sequence in (3.1) contains a subsequence whose cohomology *groups*, $H_*(M_i^n, \mathbb{Q})$, are all isomorphic. On the other hand, examples have been found showing that $\{H^*(M_i^n, \mathbb{Q})\}$ can contain infinitely many distinct *ring* structures; see [FR2].

Theorem 3.11 (π_q-Stability). ([FR2]; compare [Ro4], [Tu]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). Then for* $q \geq 2$ *and after passing to a subsequence, the* q*-th homotopy group* $\pi_q(M_i^n)$ *are all isomorphic, provided that* $\pi_q(M_i^n)$ *are finitely generated (e.g.* $sec_{M_i^n} \geq 0$ *or* $\pi_1(M_i^n)$ *is finite).*

Note that in contrast to the Betti number bound, Theorem 3.11 does not hold if upper bound on the sectional curvature is removed; see [GZ].

We now discuss sufficient topological conditions for diffeomorphism stability. Consider the sequence of fibrations, $N/\Gamma_i \to F(M_i^n) \to Y$, associated to (3.1). One would like to know when all N/Γ_i are diffeomorphic.

Proposition 3.12. ([FR4]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). If* $\pi_1(M_i^n)$ *contains no free abelian group of rank two, then* N/Γ_i *is diffeomorphic to a torus.*

In low dimensions, we have:

Theorem 3.13 (Diffeomorphism stability—low dimensions). ([FR3], [Tu]) *For* $n \leq 6$*, let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1). If* $\pi_1(M_i^n) = 1$*, then there is a subsequence all whose members are diffeomorphic.*

Note that for $n \geq 7$, one cannot expect Theorem 3.13; see [AW]. Hence, additional restrictions are required in higher dimensions. Observe that if M_i^n are 2-connected, then all $T^k \to F(M_i^n) \to Y$ are equivalent as principle T^k-bundles. In particular all $F(M_i^n)$ are diffeomorphic.

Using (3.6.2), Petrunin-Tuschmann showed that the equivalence can be chosen that is also $O(n)$-equivariant, and concluded the diffeomorphism stability for two-connected manifolds; see [PT]. For the special case in which the M_i^n are positively

pinched, the same conclusion was obtained independently in [FR1] via a different approach.

We introduce a topological condition which when M_i^n is simply connected, reduces to the assumption that $\pi_2(M_i^n)$ is finite. In the nonsimply connected case however, there are manifolds with $\pi_2(M)$ infinite, which satisfy our condition.

Let $\tilde{M}$ denote the universal covering of M. For a homomorphism, $\rho : \pi_1(M) \to \mathrm{Aut}(\mathbb{Z}^k)$, the semi-direct product, $\tilde{M} \times_{\pi_1(M)} \mathbb{Z}^k$, is a bundle of $\rho(\pi_1(M))$-modules which can be viewed as a *local coefficient system* over M. We denote it by $\mathbb{Z}^k_\rho$. Let $\tilde{b}_q(M, \mathbb{Z}^k_\rho)$ denote the rank of the cohomology group, $H^q(M, \mathbb{Z}^k_\rho)$, with the local coefficient system $\mathbb{Z}^k_\rho$. We refer to the integer

$$\tilde{b}_q(M, \mathbb{Z}^k) = \max_{\rho:\pi_1(M)\to \mathrm{Aut}(\mathbb{Z}^k)} \{\tilde{b}_q(M, \mathbb{Z}^k_\rho)\}$$

as the q-th *twisted Betti number of* M. Clearly, $\tilde{b}_q(M, \mathbb{Z}^k)$ is a topological invariant of M. Moreover, $k \cdot b_2(M, \mathbb{Z}) \leq \tilde{b}_q(M, \mathbb{Z}^k)$, with equality if $\pi_1(M) = 1$.

Theorem 3.14 (Diffeomorphism stability and geometric stability). ([FR4]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1) with* $k = n - \dim(X)$. *Assume that* M_i^n *satisfies:*
(3.14.1) $\pi_1(M_i^n)$ *is a torsion group with torsion exponents uniformly bounded from above.*
(3.14.2) The second twisted Betti number $\tilde{b}_2(M_i^n, \mathbb{Z}^k) = 0$.
Then there are diffeomorphisms, f_i, *from* M^n *to (a subsequence of)* $\{M_i^n\}$, *such that the distance functions of pullback metrics,* $f_i^*(g_i)$, *on* M^n, *converge to a pseudo-metric* d_∞ *in* C^0*-norm. Moreover,* M^n *admits a foliation with leaves diffeomorphic to flat manifolds (that are not necessarily compact) and a vector* V *tangent to a leaf if and only if* $\|V\|_{g_i} \to 0$.

The proof of Theorem 3.14 is quite involved.

Finally, we mention that J. Lott has systematically investigated the analytic aspects for a collapsing in (3.1); for details, see [Lo1-3].

4. Positively pinched manifolds

In this section, we further investigate a subclass of the class of collapsed manifolds with bounded diameter: collapsed manifolds with pinched positive sectional curvature; see [AW], [Ba], [Es], [Pü] for examples.

In the spirit of Theorem 3.4, we first give the following classification result.

Theorem 4.1 (Maximal collapse with pinched positive curvature). ([Ro8]) *Let* $M_i^n \xrightarrow{d_{GH}} X$ *be as in (3.1) such that* $sec_{M_i^n} \geq \delta > 0$. *Then* $\dim(X) \geq \frac{n-1}{2}$ *and equality implies that* $\tilde{M}_i^n \overset{\text{diffeo}}{\simeq} S^n/\mathbb{Z}_{q_i}$ *(a lens space), where* $\tilde{M}_i^n \to M_i^n$ *is a covering space of order* $\leq \frac{n+1}{2}$.

By Theorem 3.5, (3.6.1) and Proposition 3.12, the proof of Theorem 4.1 reduces to the classification of positively curved manifolds which admit invariant pure F-structures of maximal rank; see [GS].

Theorem 4.2 (Positive pinching and almost cyclicity of π_1). ([Ro6]) *Let $M_i^n \xrightarrow{d_{GH}} X$ be as in (3.1) such that $sec_{M_i^n} \geq \delta > 0$. Then for i sufficiently large, $\pi_1(M_i^n)$ has a cyclic subgroup whose index is less than $w(n)$.*

By Theorem 3.5 and (3.6.1), the following result easily implies Theorem 4.2.

Theorem 4.3 (Symmetry and almost cyclicity of π_1). ([Ro6]) *Let M^n be a closed manifold of positive sectional curvature. If M^n admits an invariant pure F-structure, then $\pi_1(M^n)$ has a cyclic subgroup whose index is less than a constant $w(n)$.*

In the special case of a free isometric action, from the homotopy exact sequence associated to the fibration, $S^1 \to M^n \to M^n/S^1$, together with the Synge theorem, one sees that $\pi_1(M^n)$ is cyclic. The proof of the general case is by induction on n and is rather complicated.

We now consider the injectivity radius estimate. Klingenberg-Sakai and Yau conjectured that the infimum of the injectivity radii of all δ-pinched metrics on M^n is a positive number which depends only on δ and the homotopy type of the manifold. By a result of Klingenberg, this conjecture is easy in even dimensions. In odd dimensions it is open.

Theorem 4.4 (Noncollapsing). ([FR4]; compare [FR1], [PT]) *For n odd, let M^n be a closed manifold satisfying $0 < \delta \leq sec_{M^n} \leq 1$ and $|\pi_1(M^n)| \leq c$. If $\tilde{b}(M^n, \mathbb{Z}^{\frac{n-1}{2}}) = 0$, then the injectivity radius of M^n is at least $\epsilon(n, \delta, c) > 0$.*

If Theorem 4.4 were false, then by Theorem 3.14 and (3.6.1) one could assume the existence of a sequence, $(M, g_i) \xrightarrow{d_{GH}} X$, with $\delta/2 \leq \sec_{g_i} \leq 1$, such that the distance functions of the metrics g_i also converge. In view of the following theorem this would lead to a contradiction.

Theorem 4.5 (Gluing). ([PRT]) *Let $(M, g_i) \xrightarrow{d_{GH}} X$ as in (1.3). If the distance functions of g_i converge to a pseudo-metric, then* $\liminf(\min sec_{g_i}) \leq 0$.

Let $f_i : (M, g_i) \xrightarrow{d_{GH}} X$ denote an ϵ_i Gromov-Hausdorff approximation, where $\epsilon_i \to 0$. For an open cover $\{B_j\}$ for X by small (contractible) balls, the assumption on the distance functions implies (roughly) that the tube, $C_{ij} = f_i^{-1}(B_j)$, is a subset of M independent of i. Clearly, the universal covering $\tilde{C}_{ij}$ of C_{ij} is noncompact. The idea is to glue together the limits of the $\tilde{C}_{ij}$ (modulo some suitable group of isometries with respect to the pullback metrics) to form a noncompact metric space with curvature bounded below by $\liminf(\min \sec_{g_i})$ in the comparison sense; see [BGP], [Pe]. On the other hand, the positivity of the curvature implies that the space so obtained would have to be compact.

The above results on δ-pinched manifolds may shed a light on the topology of positively curved manifolds. It is tempting to make the following conjecture (which seems very difficult).

Conjecture 4.6. Let M^n denote a closed manifold of positive sectional curvature.

(4.6.1) (Almost cyclicity) $\pi_1(M^n)$ has a cyclic subgroup with index bounded by a constant depending only on n.
(4.6.2) (Homotopy group finiteness) For $q \geq 2$, $\pi_q(M^n)$ has only finitely many possible isomorphism classes depending only on n and q.
(4.6.3) (Diffeomorphism finiteness) If $\pi_q(M^n) = 0$ $(q = 1, 2)$, then M^n can have only finitely many possible diffeomorphism types depending only on n.

Note that (4.6.1)–(4.6.3) are false for nonnegatively curved spaces. By the results in this section, Conjecture 4.6 would follow from an affirmative answer to the following:

Problem 4.7 (Universal pinching constant). ([Be], [Ro5]) Is there a constant $0 < \delta(n) << 1$ such that any closed n-manifold of positive sectional curvature admits a $\delta(n)$-pinched metric?

A partial verification of (4.6.2) is obtained by [FR2].

Theorem 4.8. ([FR2]) *Let M^n denote a closed manifold of positive sectional curvature. For $q \geq 2$, the minimal number of generators for $\pi_q(M^n)$ is less than $c(q, n)$.*

Previously, by Gromov the minimal number of generators of $\pi_1(M^n)$ is bounded above by a constant depending only on n.

References

[AW] S. Aloff; N. R. Wallach, *An infinite family of 7-manifolds admitting positive curved Riemannian structures*, Bull. Amer. Math. Soc. **81** (1975), 93–97.

[BGS] W. Ballmann; M. Gromov; Schroeder, *Manifolds of nonpositive curvature*, Basel:
Birkhäuser, Boston, Basel, Stuttgart, (1985).

[Ba] Ya. V. Bazaĭkin Y, *On a family of 13-dimensional closed Riemannian manifolds of positive curvature*, Sibirsk. Mat. Zh. 37 (in Russian), ii; English translation in Siberian Math. J. **6** (1996), 1068–1085.

[Be] M. Berger, *Riemannian geometry during the second half of the twentieth century*, University lecture series **17** (2000).

[BGP] Y. Burago; M. Gromov; Perel'man, *A.D. Alexandov spaces with curvature bounded below*, Uspekhi Mat. Nauk **47:2** (1992), 3–51.

[Bu1] S. Buyalo, *Collapsing manifolds of nonpositive curvature I*, Leningrad Math. J., **5** (1990), 1135–1155.

[Bu2] S. Buyalo, *Collapsing manifolds of nonpositive curvature II*, Leningrad Math. J., **6** (1990), 1371–1399.

[Bu3] S. Buyalo, *Three dimensional manifolds with Cr-structure,*, Some Questions of Geometry in the Large, A.M.S. Translations **176** (1996), 1–26.

[CCR1] J. Cao; J. Cheeger; X. Rong, *Splittings and Cr-structure for manifolds with nonpositive sectional curvature*, Invent. Math. **144** (2001), 139–167.

[CCR2] J. Cao; J. Cheeger; X. Rong, *Partial rigidity of nonpositively curved manifolds* (To appear).

[Ch] J. Cheeger, *Finiteness theorems for Riemannian manifolds*, Amer. J. Math. **92** (1970), 61–75.

[CFG] J. Cheeger; K. Fukaya; M. Gromov, *Nilpotent structures and invariant metrics on collapsed manifolds*, J. A.M.S **5** (1992), 327–372.

[CG1] J. Cheeger; M. Gromov, *Collapsing Riemannian manifolds while keeping their curvature bound I*, J. Diff. Geom **23** (1986), 309–364.

[CG2] J. Cheeger; M. Gromov, *Collapsing Riemannian manifold while keeping their curvature bounded II*, J. Differential Geom **32** (1990), 269–298.

[CG3] J. Cheeger; M. Gromov, *On the characteristic numbers of complete manifolds of bounded curvature and finite volume*, H. E. Rauch Mem Vol I (Chavel and Farkas, Eds) Springer, Berlin (1985), 115–154.

[CG4] J. Cheeger; M. Gromov, *Bounds on the von Neumann dimension of L_2-cohomology and the Gauss-Bonnet theorem for open manifolds*, J. Diff. Geom **21** (1985), 1–34.

[CR1] J. Cheeger; X. Rong, *Collapsed manifolds with bounded diameter and bounded covering geometry*, Geome.Funct. Anal **5 No. 2** (1995), 141–163.

[CR2] J. Cheeger; X. Rong, *Existence of polarized F-structure on collapsed manifolds with bounded curvature and diameter*, Geome. Funct. Anal **6, No.3** (1996), 411–429.

[Eb] P. Eberlein, *A canonical form for compact nonpositively curved manifolds whose fundamental groups have nontrivial center*, Math. Ann. **260** (1982), 23–29.

[Es] J.-H Eschenburg, *New examples of manifolds with strictly positive curvature*, Invent. Math **66** (1982), 469–480.

[FR1] F. Fang; X. Rong, *Positive pinching, volume and homotopy groups*, Geom. Funct. Anal **9** (1999), 641–674.

[FR2] F. Fang; X. Rong, *Curvature, diameter, homotopy groups and cohomology rings*, Duke Math. J. **107 No.1** (2001), 135–158.

[FR3] F. Fang; X. Rong, *Fixed point free circle actions and finiteness theorems*, Comm. Contemp. Math (2000), 75–86.

[FR4] F. Fang; X. Rong, *The twisted second Betti number and convergence of collapsing Riemannian manifolds*, To appear in Invent. Math (2002).

[Fu1] K. Fukaya, *Collapsing Riemannian manifolds to ones of lower dimension*, J. Diff. Geome **25** (1987), 139–156.

[Fu2] K. Fukaya, *Collapsing Riemannian manifolds to ones of lower dimension II*, J. Math. Soc. Japan **41** (1989), 333–356.

[Fu3] K. Fukaya, *A boundary of the set of the Riemannian manifolds with bounded curvature and diameters*, J. Diff. Geome **28** (1988), 1–21.

[Fu4] K. Fukaya, *Hausdorff convergence of Riemannian manifolds and its applications*, Recent Topics in Differential and Analytic Geometry (T. Ochiai, ed), Kinokuniya, Tokyo (1990).

[GMR] P. Granaat; M. Min-Oo; E. Ruh, *Local structure of Riemannian manifolds*, Indiana Univ. Math. J **39** (1990), 1305–1312.

[GW] D. Gromoll; J. Wolf, *Some relations between the metric structure and the algebraic structure of the fundamental group in manifolds of nonpositive*

curvature, Bull. Am. Math. Soc., **4** (1977), 545–552.

[Gr1] M. Gromov, *Almost flat manifolds*, J. Diff. Geom **13** (1978), 231–241.

[Gr2] M. Gromov, *Curvature diameter and Betti numbers*, Comment. Math. Helv **56** (1981), 179–195.

[Gr3] M. Gromov, *Volume and bounded cohomology*, I.H.E.S. Pul. Math. **56** (1983), 213–307.

[GLP] M. Gromov, J. Lafontaine; P.Pansu, *Structures metriques pour les varietes riemannienes*, CedicFernand Paris (1981).

[GK] K. Grove; H. Karcher, *How to conjugate C^1-close actions*, Math. Z **132** (1973), 11–20.

[GS] K. Grove, C. Searle, *Positively curved manifolds with maximal symmetry-rank*, J. Pure Appl. Alg **91** (1994), 137–142.

[GZ] K. Grove; W. Ziller, *Curvature and symmetry of Milnor spheres*, Ann. of Math **152** (2000), 331–367.

[LY] B. Lawson; S. T. Yau, *On compact manifolds of nonpositive curvature*, J. Diff. Geom., **7** (1972).

[Lo1] J. Lott, *Collapsing and differential form Laplacian: the case of a smooth limit space*, Duke Math. J (To appear).

[Lo2] J. Lott, *Collapsing and differential form Laplacian: the case of a smooth limit space*, Preprint (2002).

[Lo3] J. Lott, *Collapsing and Dirac-type operators*, Geometriae Delicata, Special issue on partial differential equations and their applications to geometry and physics (To appear).

[Pe] G. Perel'man, *A.D. Alexandrov spaces with curvature bounded below II*, preprint.

[Pet] P. Petersen, *Riemannian geometry*, GTM, Springer-Verlag Berlin Heidelberg New York **171** (1997).

[PRT] A. Petrunin; X. Rong; W. Tuschmann, *Collapsing vs. positive pinching*, Geom. Funct. Anal **9** (1999), 699–735.

[PT] A. Petrunin; W. Tuschmann, *Diffeomorphism finiteness, positive pinching, and second homotopy*, Geom. Funct. Anal **9** (1999).

[Pü] T. Püttmann, *Optimal pinching constants of odd dimensional homogeneous spaces*, Invent. Math **138** (1999), 631–684.

[Ro1] X. Rong, *The limiting eta invariant of collapsed 3-manifolds*, J. Diff. Geom **37** (1993), 535–568.

[Ro2] X. Rong, *The existence of polarized F-structures on volume collapsed 4-manifolds*, Geom. Funct. Anal **3, No.5** (1993), 475–502.

[Ro3] X. Rong, *Rationality of geometric signatures of complete 4-manifolds*, Invent. Math **120** (1995), 513–554.

[Ro4] X. Rong, *Bounding homotopy and homology groups by curvature and diameter*, Duke. Math. J **2** (1995), 427–435.

[Ro5] X. Rong, *On the fundamental group of manifolds of positive sectional curvature*, Ann. of Math **143** (1996), 397–411.

[Ro6] X. Rong, *The almost cyclicity of the fundamental groups of positively curved manifolds*, Invent. Math **126** (1996), 47–64.

[Ro7] X. Rong, *A Bochner Theorem and applications*, Duke Math. J **91, No.2** (1998), 381–392.

[Ro8] X. Rong, *Collapsed manifolds with pinched positive sectional curvature*, J. Diff. Geom **51 No.2** (1999), 335–358.

[Ru] E. Ruh, *Almost flat manifolds*, J. Diff. Geome. **17** (1982), 1–14.

[Sc] V. Schroeder, *Rigidity of nonpositively curved graph manifolds*, Math. Ann. **274** (1986), 19–26.

[Tu] W. Tuschmann, *Geometric diffeomorphism finiteness in low dimensions and homotopy finiteness*, Math. Ann **322** (2002), 413–420.

ICM 2002 · Vol. II · 339–349

Complex Hyperbolic Triangle Groups*

Richard Evan Schwartz†

Abstract

The theory of complex hyperbolic discrete groups is still in its childhood but promises to grow into a rich subfield of geometry. In this paper I will discuss some recent progress that has been made on complex hyperbolic deformations of the modular group and, more generally, triangle groups. These are some of the simplest nontrivial complex hyperbolic discrete groups. In particular, I will talk about my recent discovery of a closed real hyperbolic 3-manifold which appears as the manifold at infinity for a complex hyperbolic discrete group.

2000 Mathematics Subject Classification: 53.
Keywords and Phrases: Complex hyperbolic space, Discrete groups, Triangle groups, Deformations.

1. Introduction

A basic problem in geometry is the *deformation problem.* One starts with a finitely generated group Γ, a Lie group G_1, and a larger Lie group $G_2 \supset G_1$. Given a discrete embedding $\rho_0 : \Gamma \to G_1$ one asks if ρ_0 fits inside a family $\rho_t : \Gamma \to G_2$ of discrete embeddings. Here *discrete embedding* means an injective homomorphism onto a discrete set.

A nice setting for the deformation problem is the case when G_1 and G_2 are isometry groups of rank one symmetric spaces, X_1 and X_2, and Γ is isomorphic to a lattice in G_1. If $X_1 = \boldsymbol{H}^2$, the hyperbolic plane, and $X_2 = \boldsymbol{H}^3$, hyperbolic 3-space, then we are dealing with the classic and well-developed theory of quasifuchsian groups.

The (p, q, r)-*reflection triangle group* is possibly the simplest kind of lattice in $\mathrm{Isom}(\boldsymbol{H}^2)$. This group is generated by reflections in the sides of a geodesic triangle having angles π/p, π/q, π/r (subject to the inequality $1/p+1/q+1/r < 1$.) We allow the possibility that some of the integers are infinite. For instance, the $(2, 3, \infty)$-reflection triangle group is commensurable to the classical modular group.

*Supported by N.S.F. Research Grant DMS-0072706.

†Department of Mathematics, University of Maryland, College park, MD 20742, USA. E-mail: res@math.umd.edu

The reflection triangle groups are rigid in $\mathrm{Isom}(\boldsymbol{H}^3)$, in the sense that any two discrete embeddings of the same group are conjugate. We are going to replace $\boldsymbol{H}^3$ by $\boldsymbol{CH}^2$, the complex hyperbolic plane. In this case, we get nontrivial deformations. These deformations provide an attractive problem, because they furnish some of the simplest interesting examples in the still mysterious subject of complex hyperbolic deformations. While some progress has been made in understanding these examples, there is still a lot unknown about them.

In §2 we will give a rapid introduction to complex hyperbolic geometry. In §3 we will explain how to generate some complex hyperbolic triangle groups. In §4 we will survey some results about these groups and in §5 we will present a more complete conjectural picture. In §6 we will indicate some of the techniques we used in proving our results.

2. The complex hyperbolic plane

The book [8] is an excellent general reference for complex hyperbolic geometry. Here are some of the basics.

$\boldsymbol{C}^{2,1}$ is a copy of the vector space $\boldsymbol{C}^3$ equipped with the Hermitian form

$$\langle U, V\rangle = -u_3\overline{v}_3 + \sum_{j=1}^{n} u_j\overline{v}_j. \tag{1}$$

Here $U = (u_1, u_2, u_3)$ and $V = (v_1, v_2, v_3)$. A vector V is called *negative*, *null*, or *positive* depending (in the obvious way) on the sign of $\langle V, V\rangle$. We denote the set of negative, null, and positive vectors, by N_-, N_0 and N_+ respectively.

$\boldsymbol{C}^2$ includes in complex projective space $\boldsymbol{CP}^2$ as the affine patch of vectors with nonzero last coordinate. Let $[\,] : \boldsymbol{C}^{2,1} - \{0\} \to \boldsymbol{CP}^2$ be the projectivization whose formula, expressed in the affine patch, is

$$[(v_1, v_2, v_3)] = (v_1/v_3, v_2/v_3). \tag{2}$$

The *complex hyperbolic plane*, $\boldsymbol{CH}^2$, is the projective image of the set of negative vectors in $\boldsymbol{C}^{2,1}$. That is, $\boldsymbol{CH}^2 = [N_-]$. The ideal boundary of $\boldsymbol{CH}^2$ is the unit sphere $S^3 = [N_0]$. If $[X], [Y] \in \boldsymbol{CH}^n$ the complex hyperbolic distance $\varrho([X], [Y])$ satisfies

$$\varrho([X], [Y]) = 2\cosh^{-1}\sqrt{\delta(X, Y)}; \qquad \delta(X, Y) = \frac{\langle X, Y\rangle\langle Y, X\rangle}{\langle X, X\rangle\langle Y, Y\rangle}. \tag{3}$$

Here X and Y are arbitrary lifts of $[X]$ and $[Y]$. See [8, 77]. The distance we defined is induced by an invariant Riemannian metric of sectional curvature pinched between -1 and -4. This Riemannian metric is the real part of a Kähler metric.

$SU(2, 1)$ is the Lie group of $\langle\ ,\ \rangle$ preserving complex linear transformations. $PU(2, 1)$ is the projectivization of $SU(2, 1)$ and acts isometrically on $\boldsymbol{CH}^2$. The map $SU(2, 1) \to PU(2, 1)$ is a 3-to-1 Lie group homomorphism. The group of holomorphic isometries of $\boldsymbol{CH}^2$ is exactly $PU(2, 1)$. The full group of isometries

of $\boldsymbol{CH}^2$ is generated by $PU(2,1)$ and by the antiholomorphic map $(z_1, z_2, z_3) \to (\overline{z}_1, \overline{z}_2, \overline{z}_3)$.

An element of $PU(2,1)$ is called *elliptic* if it has a fixed point in $\boldsymbol{CH}^2$. It is called *hyperbolic* (or *loxodromic*) if there is some $\epsilon > 0$ such that every point in $\boldsymbol{CH}^2$ is moved at least ϵ by the isometry. An element which is neither elliptic nor hyperbolic is called *parabolic.*

$\boldsymbol{CH}^2$ has two different kinds of totally geodesic subspaces, *real slices* and *complex slices.* Every real slice is isometric to $\boldsymbol{CH}^2 \cap \boldsymbol{R}^2$ and every complex slice is isometric to $\boldsymbol{CH}^2 \cap \boldsymbol{C}^1$. The ideal boundaries of real and complex slices are called, respectively, $\boldsymbol{R}$-circles and $\boldsymbol{C}$-circles. The complex slices naturally implement the Poincaré model of the hyperbolic plane and the real slices naturally model the Klein model. It is a beautiful feature of the complex hyperbolic plane that it contains both models of the hyperbolic plane.

3. Reflection triangle groups

There are two kinds of reflections in $\mathrm{Isom}(\boldsymbol{CH}^2)$. A *real reflection* is an anti-holomorphic isometry conjugate to the map $(z, w) \to (\overline{z}, \overline{w})$. The fixed point set of a real reflection is a real slice. We shall not have much to say about the explicit computation of real reflections, but rather will concentrate on the complex reflections.

A *complex reflection* is a holomorphic isometry conjugate to the involution $(z, w) \to (z, -w)$. The fixed point set of a complex reflection is a complex slice. There is a simple formula for the general complex reflection: Let $C \in N_+$. Given any $U \in \boldsymbol{C}^{2,1}$ define

$$I_C(U) = -U + \frac{2\langle U, C\rangle}{\langle C, C\rangle} C. \tag{4}$$

I_C is a complex reflection.

We also have the formula

$$U \boxtimes V = (\overline{u_3 v_2 - u_2 v_3},\ \overline{u_1 v_3 - u_3 v_1},\ \overline{u_1 v_2 - u_2 v_1}). \tag{5}$$

This vector is such that $\langle U, U \boxtimes V\rangle = \langle V, U \boxtimes V\rangle = 0$. See [8, p. 45].

Equations 4 and 5 can be used in tandem to rapidly generate triangle groups defined by complex reflections. One picks three vectors $V_1, V_2, V_2 \in N_-$. Next, we let $C_j = V_{j-1} \boxtimes V_{j+1}$. Indices are taken mod 3. Finally, we let $I_j = I_{C_j}$. The complex reflection I_j fixes the complex line determined by the points $[V_{j-1}]$ and $[V_{j+1}]$. This, the group $\langle I_1, I_2, I_3\rangle$ is a complex-reflection triangle group determined by the triangle with vertices $[V_1]$, $[V_2]$, $[V_3]$.

Here is a quick dimension count for the space of (p, q, r)-triangle groups generated by complex reflections. We can normalize so that $[V_1] = 0$. The stabilizer of 0 in $PU(2,1)$ acts transitively on the unit tangent space at 0. We can therefore normalize so that $[V_2] = (s, 0)$ where $s \in (0, 1)$. Finally, the isometries $(z, w) \to (z, \exp(i\theta)w)$ stabilize both $[V_1]$ and $[V_2]$. Applying a suitable isometry we arrange that $[V_3] = (t+iu, v)$ where $t, u, v \in (0, 1)$. We cannot make any further normalizations, so the space of triangles in $\boldsymbol{CH}^2$ mod isometry is 4-real dimensional.

Each of the three angles (p, q, r) puts 1 real constraint on the triangle. For instance, the p-angle places the constraint that $(I_1 I_2)^p$ is the identity. Since $4 - 3 = 1$, we see heuristically that the space of (p, q, r)-complex reflection triangle groups is 1-real dimensional.

The argument we just gave can be made rigorous, and extends to the case when some of the integers are infinite. (In this case the corresponding vectors are null rather than negative.) In the (∞, ∞, ∞)-case, the parameter is the *angular invariant* $\arg(\langle V_1, V_2\rangle\langle V_2, V_3\rangle\langle V_3, V_1\rangle)$. Compare [10].

This 1-dimensionality of the deformation space makes the (p, q, r)-triangle groups an especially attractive problem to study. Indeed, there is a completely canonical path of deformations. The starting point for the path of deformations is the case when the vectors have entirely real entries. (That is, $u = 0$.) In this case, the three complex reflections stabilize the real slice $\boldsymbol{R}^2 \cap \boldsymbol{CH}^2$.

4. Some results

To obtain a deformation of the (p, q, r)-reflection triangle group we choose a slice, either real or complex, and a triple of reflections, either real or complex, which restrict to the reflections in the sides of a (p, q, r)-geodesic triangle in the slice. *A priori* there are 4 possibilities, given that the slice and the reflection types can be either real or complex. These choices lead to different outcomes.

If we start with complex reflections stabilizing a complex slice, the group has order 2, because the reflections will all stabilize the same slice.

A more interesting case involving complex slices is given by:

Theorem 4.1 [8] *$\rho_0 : \Gamma \to Isom(\boldsymbol{CH}^2)$ stabilizes a complex slice and acts on this slice with compact quotient then any nearby representation ρ_t also stabilizes a complex slice.*

Goldman's theorem applies to any co-compact lattice $\rho_0(\Gamma)$. In the case of triangle groups, which are rigid in $\boldsymbol{H}^2$, it says that any nearby representation is conjugate the original. In contrast:

Theorem 4.2 [4, 12] *There is a 1-parameter family $\rho_t(\Gamma(2, 3, \infty))$ of discrete faithful representations of the modular group having the property that ρ_0 stabilizes a real slice and ρ_1 stabilizes a complex slice. For every parameter the generators are real reflections.*

Thus, in the case of non-cocompact triangle groups, two of the remaining 3 cases can be *connected*. In their paper, Falbel and Koseleff claim that their technique works for $\Gamma(p, q, \infty)$ when $\max(p, q) = 4$. For higher values of p and q it is not known what happens.

The remaining case occurs when we start with complex reflections stabilizing a real slice. This is the case we discussed in the previous section. Henceforth we restrict our attention to this case.

Goldman and Parker introduced this topic and studied the case of the ideal triangle group $\Gamma(\infty, \infty, \infty)$. They found that there is a 1-real parameter family of

non-conjugate representations, $\{\rho_t,\ t \in (-\infty, \infty)\}$. Once again ρ_0 stabilizes a real slice. Paraphrasing their more precise formulation:

Theorem 4.3 [10] *There are symmetric neighborhoods $I \subset J$ of 0 such that ρ_t is discrete and faithful if $t \in I$ and not both discrete and faithful if $t \notin J$.*

J consists of the parameter values t such that the element $\rho_t(I_1 I_2 I_3)$ is not an elliptic element. For $t \notin J$, this element is elliptic. If it has finite order then the representation is not faithful; if it has infinite order then the representation is not discrete. The (very slightly) smaller interval I is the interval for which their proof works. They conjectured that ρ_t should be discrete and faithful iff $t \in J$.

We proved the Goldman-Parker conjecture, and sharpened it a bit.

Theorem 4.4 [16] *ρ_t is discrete and faithful if and only if $t \in J$. Furthermore, ρ_t is indiscrete if $t \notin J$.*

The group $L = \rho_s(\Gamma(\infty, \infty, \infty))$, when $s \in \partial J$ is especially beautiful. We call this group the *last ideal triangle group.* (There are really two groups, one for each endpoint of J, but these are conjugate.) This group seems central in the study of complex hyperbolic deformations of the modular group. For instance, Falbel and Parker recently discovered that L arises as the endpoint of a certain family of deformations of the modular group, using real reflections. See [5] for details.

Recall that L, like all discrete groups, has a *limit set* $\Omega(L) \subset S^3$ and a *domain of discontinuity* $\Delta(L) = S^3 - \Omega(L)$. The quotient $\Delta(L)/L$ is a 3-dimensional orbifold, commonly called the *orbifold at infinity.*

Theorem 4.5 [17] *$\Delta(L)/L$ is commensurable to the Whitehead link complement.*

The Whitehead link complement is a classic example of a finite volume hyperbolic 3-manifold. The surprise in the above result is that a real hyperbolic 3-manifold makes its appearance in the context of complex hyperbolic geometry.

One might wonder about analogues of Theorem 4.4 for other triangle groups. Below we will conjecture that the space of discrete embeddings is a certain interval. In his thesis [22], Justin Wyss-Gallifent studied some special cases of this question. He made a very interesting discovery concerning the $(4, 4, \infty)$ triangle group:

Theorem 4.6 [22] *Let S be the set of parameters t for which the representation $\rho_t(\Gamma(4, 4, \infty))$ is discrete (but not necessarily injective). Then S contains isolated points and, in particular, is not an interval.*

There seems to be an interval J of discrete embeddings and, outside of J, an extra countable sequence $\{t_j\}$ of parameters for which ρ_{t_j} is discrete but not an embedding. This sequence accumulates on the endpoints of J.

Motivated by [17] I wanted to produce a discrete complex hyperbolic group whose orbifold at infinity was a closed hyperbolic 3-manifold. The extra representations found by Wyss-Gallifent seemed like a good place to start. Unfortunately, there is a cusp built into the representations of the $(4, 4, \infty)$ triangle groups.

Instead, I considered the $(4,4,4)$-groups, and found that the extra discrete deformations exist. $\rho_t(\Gamma(4,4,4))$ seems to be discrete embedding iff all the elements of the form $\rho_t(I_iI_jI_iI_k)$ are not elliptic. Here i,j,k are meant to be distinct. (For all these parameters, the element $\rho_t(I_iI_jI_k)$ is still a loxodromic element.) There is a countable collection $t_5, t_6, ...$ of parameters such that $\rho_{t_j}(I_iI_jI_iI_k)$ has order j. All these representations seem discrete. For ease of notation we set $\rho_j = \rho_{t_j}$.

For $j = 5,6,7,8,12$ we can show by arithmetic means that ρ_j is discrete. The representation ρ_5 was too complicated for me to analyze and ρ_6 has a cusp. The simplest remaining candidate is ρ_7.

Theorem 4.7 [18] *$G = \rho_7(\Gamma(4,4,4))$ is a discrete group. The orbifold at infinity $\Delta(G)/G$ is a closed hyperbolic 3-orbifold.*

In the standard terminology, $\Delta(G)/G$ is the orbifold obtained by labelling the braid $(AB)^{15}(AB^{-2})^3$ with a 2. Here A and B are the standard generators of the 3-strand braid group.

A *spherical CR structure* on a 3-manifold is a system of coordinate charts into S^3 whose transition functions are restrictions of complex projective transformations. Kamishima and Tsuboi [13] produced examples of spherical CR structures on Seifert fibered 3-manifolds, but our example in theorem 4.7 gives the only known spherical CR structure on a closed hyperbolic 3-manifold. We think that Theorem 4.7 holds for all $j = 8,9,10....$

Concerning the specific topic of triangle groups generated by complex reflections, I think that not much else is known. Recently a lot of progress has been made in understanding triangle groups generated by real reflections. See [3] and [4]. There has been a lot of other great work done recently on complex hyperbolic discrete groups, for instance [1], [2], [9], [20], [21]. Also see the references in Goldman's book [f8].

5. A conjectural picture

We will consider the 1-parameter family $\rho_t(p,q,r)$ of representations of the (p,q,r)-reflection triangle group, using complex reflections. We arrange that ρ_0 stabilizes a real slice. We choose our integers so that $p \le q \le r$. We let I_p, I_q, I_r be the generators of the reflection triangle group. The notation is such that I_p is the reflection in the side of the triangle opposite p, etc. Define

$$W_A = I_pI_rI_qI_r; \qquad W_B = I_pI_qI_r. \tag{6}$$

Conjecture 5.1 *The set of t for which $\rho_t(p,q,r)$ is a discrete embedding is the closed interval consisting of the parameters t for which neither $\rho_t(W_A)$ nor $\rho_t(W_B)$ is elliptic.*

We call the interval of Conjecture 5.1 the *critical interval.*

We say that the triple (p,q,r) has *type A* if the endpoints of the critical interval correspond to the representations when W_A is a parabolic element. In other words, W_A becomes elliptic before W_B. We say otherwise that (p,q,r) has *type B.*

Conjecture 5.2 *The triple* (p,q,r) *has type A if* $p < 10$ *and type B if* $p > 13$.

The situation is rather complicated when $p \in \{10,11,12,13\}$. Our Java applet [19] lets the user probe these cases by hand, though the roundoff error makes a few cases ambiguous. The extra deformation, which was the subject of Theorem 4.7, seems part of a more general pattern.

Conjecture 5.3 *If* (p,q,r) *has type A then there is a countable collection of parameters* $t_1, t_2, t_3 \ldots$ *for which* $\rho_{t_j}(p,q,r)$ *is infinite and discrete but not injective. If* (p,q,r) *has type B then all infinite discrete representations* $\rho_t(p,q,r)$ *are embeddings and covered by Conjecture 5.1.*

The *proviso* about the infinite image arises because there always exists an extremely degenerate representation of $\Gamma(p,q,r)$ onto $\boldsymbol{Z}/2$. The generators are all mapped to the same complex reflection.

In summary, there seems to be a critical interval I, such the representations $\rho_t(p,q,r)$ are discrete embeddings iff $t \in I$. Depending on the endpoints of I, there are either no additional discrete representations, or a countable collection of extra discrete representations.

It is interesting to see what happens as t moves to the boundary of I from within I. We observed a certain kind of monotonicity to the way the representation varies. Let Γ be the abstract (p,q,r) triangle group. For any word $W \in \Gamma$, let $W_t = \rho_t(W)$. We will concentrate on the case when W is an infinite word. For $t \in I$, the element W_t is (conjecturally) either a parabolic or loxodromic. Let $\lambda(W_t)$ be the translation length of W_t.

Conjecture 5.4 *As* t *increases monotonically from* 0 *to* ∂I, *the quantity* $\lambda(W_t)$ *decreases monotonically for all infinite words* W.

Conjecture 5.4 is closely related to some conjectures of Hanna Sandler [15] about the behavior of the trace function in the ideal triangle case. I think that there is some fascinating algebra hiding behind the triangle groups—in the form of the behavior of the trace function—but so far it is unreachable.

6. Some techniques of proof

If $G \subset \mathrm{Isom}(X)$, one can try to show that G is discrete by constructing a *fundamental domain* for G. One looks for a set $F \subset X$ such that the orbit $G(F)$ *tiles* X. This means that the translates of F only intersect F in its boundary. The Poincaré theorem [B, §9.6] gives a general method for establishing the tiling property of F based on how certain elements of G act on ∂F.

When $X = \boldsymbol{H}^n$, one typically builds fundamental domains out of polyhedra bounded by totally geodesic codimension-1 faces. When $X = \boldsymbol{CH}^n$, the situation is complicated by the absence of totally geodesic codimension-1 subspaces. The most natural replacement is the *bisector*. A bisector is the set of points in $\boldsymbol{CH}^n$ equidistant between two given points. Mostow [14] used bisectors in his analysis

of some exceptional non-arithmetic lattices in $\mathrm{Isom}(\boldsymbol{CH}^2)$, and Goldman studied them extensively in [8]. (See Goldman's book for additional references on papers which use bisectors to construct fundamental domains.)

My point of view is that there does not seem to be a "best" kind surface to use in constructing fundamental domains in complex hyperbolic space. Rather, I think that one should be ready to fabricate new kinds of surfaces to fit the problem at hand. It seems that computer experimentation often reveals a good choice of surface to use. In what follows I will give a quick tour of constructive techniques.

Consider first the deformations $G_t = \rho_t(\infty, \infty, \infty)$ of the ideal triangle group, introduced in [10]. According to [16] these groups are discrete for $t \in [0, \tau]$. Here τ is the *critical parameter* where the product of the generators is parabolic. It is convenient to introduce the *Clifford torus*. Thinking of $\boldsymbol{CH}^2$ as the open unit ball in $\boldsymbol{C}^2$, the Clifford torus is the subset $T = \{|z| = |w|\} \subset S^3$. Amazingly T has 3 foliations by $\boldsymbol{C}$-circles: The *horizontal foliation* consists of $\boldsymbol{C}$-circles of the form $\{(z,w)|z = z_0\}$. The *vertical foliation* consists of $\boldsymbol{C}$-circles of the form $\{(z,w)|w = w_0\}$. The *diagonal foliation* consists of $\boldsymbol{C}$-circles having the form $\{(z,w)|\ z = \lambda_0 w\}$.

Recall that G_t is generated by 3 complex reflections. Each of these reflections fixes a complex slice and hence the bounding $\boldsymbol{C}$-circle. One can normalize so that the three fixed $\boldsymbol{C}$-circles lie on the Clifford torus, one in each of the foliations. Passing to an index 2 subgroup, we can consider a group generated by 4 complex reflections: Two of these reflections, H_1 and H_2, fix horizontal $\boldsymbol{C}$-circles h_1 and h_2 and the other two, V_1 and V_2, fix vertical $\boldsymbol{C}$-circles v_1 and v_2.

The ideal boundary of a bisector is called a *spinal sphere*. This is an embedded 2-sphere which is foliated by $\boldsymbol{C}$-circles (and also by $\boldsymbol{R}$-circles.) We can find a configuration of 4-spinal spheres $S(1,v)$, $S(2,v)$, $S(1,h)$ and $S(2,h)$. Here $S(j,v)$ contains v_j as part of its foliation and $S(j,h)$ contains h_j as part of its foliation. The map H_j stabilizes $S(j,h)$ and interchanges the two components of $S^3 - S(j,h)$. Analogous statements apply to the Vs.

The two spheres $S(h,j)$ are contained in the closure of one component of $S^3 - T$ and the two spheres $S(v,j)$ are contained in the closure of the other. When the parameter t is close to 0 these spinal spheres are all disjoint from each other, excepting tangencies, and form a kind of necklace of spheres. Given the way the elements H_j and V_j act on our necklace of spheres, we see that we are dealing with the usual picture associated to a *Schottky group*. In this case the discreteness of the group is obvious.

As the parameter increases, the two spinal spheres $S(v,1)$ and $S(v,2)$ collide. Likewise, $S(h,1)$ and $S(h,2)$ collide. Unfortunately, the collision parameter occurs before the critical parameter. For parameters larger than this collision parameter, we throw out the spinal spheres and look at the action of G on the Clifford torus itself. (This is not the point of view taken in [10] but it is equivalent to what they did.)

Let H be the subgroup generated by the reflections H_1 and H_2. One finds that the orbit $H(T)$ consists of translates of T which are disjoint from each other except for forced tangencies. Even though H is an infinite group, most of the elements in

H move T well off itself, and one only needs to take care in checking a short finite list of words in H. Once we know how H acts on T we invoke a variant of the *ping-pong lemma* to get the discreteness.

At some new collision parameter, the translates of the Clifford torus collide with each other. Again, the collision parameter occurs before the critical parameter. This is where the work in [16] comes in. I define a new kind of surface called a *hybrid cone*. A hybrid cone is a certain surface foliated by arcs of $\boldsymbol{R}$-circles. These arcs make the pattern of a fan: Each arc has one endpoint on the arc of a $\boldsymbol{C}$-circle and the other endpoint at a single point common to all the arcs. I cut out two triangular patches on the Clifford torus and replace each patch by a union of three hybrid cones. Each triangular patch is bounded by three arcs of $\boldsymbol{C}$-circles; so that the hybrid cones are formed by connecting these exposed arcs to auxilliary points using arcs of $\boldsymbol{R}$-circles. In short, I put some dents into the Clifford torus to make it fit better with its H-translates, and the I apply the ping-pong lemma to the dented torus.

I also use hybrid cones in [17], to construct a natural fundamental domain in the domain of discontinuity $\Delta(L)$ for the last ideal triangle group L. In this case, the surfaces fit together to make three topological spheres, each tangent to the other two along arcs of $\boldsymbol{R}$-circles. The existence of this fundamental domain lets me compute explicitly that $\Delta(L)/L$ is commensurable to the Whitehead link complement.

Falbel and Zocca [6] introduce related surfaces called $\boldsymbol{C}$-spheres, which are foliated by $\boldsymbol{C}$-circles. These surfaces seem especially well adapted to groups generated by real reflections. See [3] and [4]. Indeed, Falbel and Parker construct a different fundamental domain for L using $\boldsymbol{C}$-spheres. See [5].

To prove Theorem 4.7 in [18] I introduce another method of constructing fundamental domains. My proof revolves around the construction of a simplicial complex $Z \subset \boldsymbol{C}^{2,1}$. The vertices of Z are canonical lifts to $\boldsymbol{C}^{2,1}$ of fixed points of certain elements of the group $G = \rho_7(\Gamma(4,4,4))$. The tetrahedra of Z are Euclidean convex hulls of various 4-element subsets of the vertices. Comprised of infinitely many tetrahedra, Z is invariant under the element $I_2 I_1 I_3$. Modulo this element Z has only finitely many tetrahedra.

Recall that $[\ ]$ is the projectivization map. Let $[Z_0] = [Z] \cap S^3$. I deduce the topology of the orbifold at infinity by studying the topology of $[Z_0]$. To show that my analysis of the topology at infinity is correct, I show that one component F of $\boldsymbol{CH}^2 - [Z]$ has the *tiling property*: The G-orbit of F tiles $\boldsymbol{CH}^2$. Now, Z is an essentially combinatorial object, and it not too hard to analyze the combinatorics and topology of Z in the abstract. The hard part is showing that the map $Z \to [Z]$ is an embedding. Assuming the embedding, the combinatorics and topology of Z are reproduced faithfully in $[Z]$, and I invoke a variant of the Poincaré theorem.

After making some easy estimates, my main task boils down to showing that the projectivization map $[\ ]$ is injective on all pairs of tetrahedra within a large but finite portion of Z. Roughly, I need to check about 1.3 million tetrahedra. The sheer number of checks forces us to bring in the computer. I develop a technique for proving, with rigorous machine-aided computation, that $[\ ]$ is injective on a given

pair of tetrahedra.

A novel feature of my work is the use of computer experimentation and computer-aided proofs. This feature is also a drawback, because it only allows for the analysis of examples one at a time. To make this analysis automatic I would like to see a kind of marriage of complex hyperbolic geometry and computation. On the other hand, I would greatly prefer to see some theoretical advances in discreteness-proving which would eliminate the computer entirely.

References

[1] D. Allcock, J. Carlson, D. Toledo, *The Moduli Space of Cubic Threefolds*, J. Alg. Geom. (to appear).

[2] P. Deligne and G.D. Mostow, *Commensurabilities among Lattices in $PU(1,n)$*, Annals of Mathematics Studies **132**, Princeton University Press (1993).

[3] E. Falbel and P.-V. Koseleff, *Flexibility of the Ideal Triangle Group in Complex Hyperbolic Geometry*, Topology **39(6)** (2000), 1209–1223.

[4] E. Falbel and P.-V. Koseleff, *A Circle of Modular Groups*, preprint 2001.

[5] E. Falbel and J. Parker, *The Moduli Space of the Modular Group in Complex Hyperbolic Geometry*, Math. Research Letters (to appear).

[6] E. Falbel and V. Zocca, *A Poincaré's Fundamental Polyhedron Theorem for Complex Hyperbolic Manifolds*, J. reine angew Math. **516** (1999), 133–158.

[7] W. Goldman *Representations of fundamental groups of surfaces*, in "Geometry and Topology, Proceedings, University of Maryland 1983–1984", J. Alexander and J. Harer (eds.), Lecture Notes in Math. Vol. 1167 (1985), 95–117.

[8] W. Goldman, *Complex Hyperbolic Geometry*, Oxford Mathematical Monographs, Oxford University Press, (1999).

[9] W. Goldman, M. Kapovich and B. Leeb, *Complex Hyperbolic Surfaces Homotopy Equivalent to a Riemann surface*, Communications in Analysis and Geometry **9** (2001), 61–95.

[10] W. Goldman and J. Parker, *Complex Hyperbolic Ideal Triangle Groups*, J. reine agnew Math. **425** (1992), 71–86.

[11] W. Goldman and J. Millson, *Local Rigidity of Discrete Groups Acting on Complex Hyperbolic Space*, Inventiones Mathematicae **88** (1987), 495–520.

[12] N. Gusevskii and J.R. Parker, *Complex Hyperbolic Representations of Surface Groups and Toledo's Invariant*, preprint (2001).

[13] Y. Kamishima and T. Tsuboi, *CR Structures on Seifert Manifolds*, Invent. Math. **104** (1991) 149–163.

[14] G.D. Mostow, *On a Remarkable Class of Polyhedra in Complex Hyperbolic Space*, Pac. Journal of Math **86** (1980) 171–276.

[15] H. Sandler, *Trace Equivalence in $SU(2,1)$*, Geo Dedicata **69** (1998) 317–327.

[16] R. E. Schwartz, *Ideal Triangle Groups, Dented Tori, and Numerical Analysis*, Annals of Math **153** (2001).

[17] R.E. Schwartz, *Degenerating the Complex Hyperbolic Ideal Triangle Groups*, Acta Mathematica **186** (2001).

[18] R. E. Schwartz, *Real Hyperbolic on the Outside, Complex Hyperbolic on the*

Inside, Invent. Math (to apear).
[19] R. E. Schwartz, *Applet 29* (2001) http://www.math.umd.edu/~res.
[20] Y. Shalom, *Rigidity, Unitary Representations of Semisimple Groups, and Fundamental Groups of Manifolds with Rank One Transformation Group*, Annals of Math **152** (2000) 113–182.
[21] D. Toledo, *Representations of Surface Groups on Complex Hyperbolic Space*, Journal of Differential Geometry **29** (1989) 125–133.
[22] J. Wyss-Gallifent, *Discreteness and Indiscreteness Results for Complex Hyperbolic Triangle Groups*, Ph.D. Thesis, University of Maryland (2000).

ICM 2002 · Vol. II · 351–360

Fukaya Categories and Deformations

Paul Seidel*

Abstract

It is widely believed that the right "cycles" for symplectic geometry are Lagrangian submanifolds of symplectic manifolds (see for instance Weinstein's 1981 survey). This can be given several different meanings, depending on the kind of symplectic geometry one is interested in. In one direction, the development of Floer cohomology for Lagrangian submanifolds, culminating in recent work of Fukaya, Oh, Ohta and Ono, has led to the definition of a "Fukaya category" associated to a symplectic manifold. I want to look at the relation between the Fukaya category of an affine variety $M \subset \mathbb{C}^N$ and that of its projective closure $\bar{M} \subset \mathbb{C}P^N$. This can be set up as a "deformation problem" in the abstract algebraic sense.

2000 Mathematics Subject Classification: 57R17, 57R56, 18E30.

Soon after their first appearance [7], Fukaya categories were brought to the attention of a wider audience through the homological mirror conjecture [14]. Since then Fukaya and his collaborators have undertaken the vast project of laying down the foundations, and as a result a fully general definition is available [9, 6]. The task that symplectic geometers are now facing is to make these categories into an effective tool, which in particular means developing more ways of doing computations in and with them.

For concreteness, the discussion here is limited to projective varieties which are Calabi-Yau (most of it could be carried out in much greater generality, in particular the integrability assumption on the complex structure plays no real role). The first step will be to remove a hyperplane section from the variety. This makes the symplectic form exact, which simplifies the pseudo-holomorphic map theory considerably. Moreover, as far as Fukaya categories are concerned, the affine piece can be considered as a first approximation to the projective variety. This is a fairly obvious idea, even though its proper formulation requires some algebraic formalism of deformation theory. A basic question is the finite-dimensionality of the relevant

*Centre de Mathematiques, Ecole Polytechnique, F-91128 Palaiseau, France. E-mail: seidel@math.polytechnique.fr

deformation spaces. As Conjecture 4 shows, we hope for a favourable answer in many cases. It remains to be seen whether this is really a viable strategy for understanding Fukaya categories in interesting examples.

Lack of space and ignorance keeps us from trying to survey related developments, but we want to give at least a few indications. The idea of working relative to a divisor is very common in symplectic geometry; some papers whose viewpoint is close to ours are [12, 16, 3, 17]. There is also at least one entirely different approach to Fukaya categories, using Lagrangian fibrations and Morse theory [8, 15, 4]. Finally, the example of the two-torus has been studied extensively [18].

1. Symplectic cohomology

We will mostly work in the following setup:

Assumption 1. *X is a smooth complex projective variety with trivial canonical bundle, and D a smooth hyperplane section in it. We take a suitable small open neighbourhood $U \supset D$, and consider its complement $M = X \setminus U$. Both X and M are equipped with the restriction of the Fubini-Study Kähler form. Then M is a compact exact symplectic manifold with contact type boundary, satisfying $c_1(M) = 0$.*

Consider a holomorphic map $u : \Sigma \to X$, where Σ is a closed Riemann surface. The symplectic area of u is equal (up to a constant) to its intersection number with D. When counting such maps in the sense of Gromov-Witten theory, it is convenient to arrange them in a power series in one variable t, where the t^k term encodes the information from curves having intersection number k with D. The t^0 term corresponds to constant maps, hence is sensitive only to the classical topology of X. Thus, for instance, the small quantum cohomology ring $QH^*(X)$ is a deformation of the ordinary cohomology $H^*(X)$.

As we've seen, there are only constant holomorphic maps from closed Riemann surfaces to $M = X \setminus D$. But one can get a nontrivial theory by using punctured surfaces, and deforming the holomorphic map equation near the punctures through an inhomogeneous term, which brings the Reeb dynamics on ∂M into play. This can be done more generally for any exact symplectic manifold with contact type boundary, and it leads to the symplectic cohomology $SH^*(M)$ of Cieliebak-Floer-Hofer [2] and Viterbo [26, 27]. Informally one can think of $SH^*(M)$ as the Floer cohomology $HF^*(M \setminus \partial M, H)$ for a Hamiltonian function H on the interior whose gradient points outwards near the boundary, and becomes infinite as we approach the boundary. For technical reasons, in the actual definition one takes the direct limit over a class of functions with slower growth (to clarify the conventions: our $SH^k(M)$ is dual to the $FH^{2n-k}(M)$ in [26]). The algebraic structure of symplectic cohomology is different from the familiar case of closed M, where one has large quantum cohomology and the WDVV equation. Operations $SH^*(M)^{\otimes p} \to SH^*(M)^{\otimes q}$, for $p \geq 0$ and $q > 0$, come from families of Riemann surfaces with $p + q$ punctures, together with a choice of local coordinate around each puncture. The Riemann surfaces may degenerate to stable singular ones, but only if no component of the

normalization contains some of the first p and none of the last q punctures. This means that if we take only genus zero and $q = 1$ then no degenerations at all are allowed, and the resulting structure is that of a Batalin-Vilkovisky (BV) algebra [10]. For instance, let $M = D(T^*L)$ be a unit cotangent bundle of an oriented closed manifold L. Viterbo [27] computed that $SH^*(M) \cong H_{n-*}(\Lambda L)$ is the homology of the free loop space, and a reasonable conjecture says that the BV structure agrees with that of Chas-Sullivan [1].

Returning to the specific situation of Assumption 1, and supposing that U has been chosen in such a way that the Reeb flow on ∂M becomes periodic, one can use a Bott-Morse argument [19] to get a spectral sequence which converges to $SH^*(M)$. The starting term is

$$E_1^{pq} = \begin{cases} H^q(M) & p = 0, \\ H^{q+3p}(\partial M) & p < 0. \end{cases} \tag{1}$$

It might be worth while to investigate this further, in order to identify the differentials (very likely, a version of the relative Gromov-Witten invariants [12] for $D \subset X$). But even without any more effort, one can conclude that each group $SH^k(M)$ is finite-dimensional. In particular, assuming that $dim_{\mathbb{C}}(X) > 2$ (and appealing to hard Lefschetz, which will be the only time that we use any algebraic geometry) one has

$$dim\, SH^2(M) \leq b_2(M) + b_0(\partial M) = b_2(X). \tag{2}$$

2. Fukaya categories

M (taken as in Assumption 1) is an exact symplectic manifold, and there is a well-defined notion of exact Lagrangian submanifold in it. Such submanifolds L have the property that there are no non-constant holomorphic maps $u : (\Sigma, \partial\Sigma) \to (M, L)$ for a compact Riemann surface Σ, hence a theory of "Gromov-Witten invariants with Lagrangian boundary conditions" would be trivial in this case. To get something interesting, one removes some boundary points from Σ, thus dividing the boundary into several components, and assigns different L to them. The part of this theory where Σ is a disk gives rise to the Fukaya A_∞-category $\mathcal{F}(M)$.

The basic algebraic notion is as follows. An A_∞-category $\mathcal{A}$ (over some field, let's say $\mathbb{Q}$) consists of a set of objects $Ob\mathcal{A}$, and for any two objects a graded $\mathbb{Q}$-vector space of morphisms $hom_{\mathcal{A}}(X_0, X_1)$, together with composition operations

$$\begin{aligned} &\mu^1_{\mathcal{A}} : hom_{\mathcal{A}}(X_0, X_1) \longrightarrow hom_{\mathcal{A}}(X_0, X_1)[1], \\ &\mu^2_{\mathcal{A}} : hom_{\mathcal{A}}(X_1, X_2) \otimes hom_{\mathcal{A}}(X_0, X_1) \longrightarrow hom_{\mathcal{A}}(X_0, X_2), \\ &\mu^3_{\mathcal{A}} : hom_{\mathcal{A}}(X_2, X_3) \otimes hom_{\mathcal{A}}(X_1, X_2) \otimes hom_{\mathcal{A}}(X_0, X_1) \to \\ &\qquad \longrightarrow hom_{\mathcal{A}}(X_0, X_3)[-1], \quad \ldots. \end{aligned}$$

These must satisfy a sequence of quadratic "associativity" equations, which ensure that $\mu^1_{\mathcal{A}}$ is a differential, $\mu^2_{\mathcal{A}}$ a morphism of chain complexes, and so on. Note that by

forgetting all the $\mu^d_{\mathcal{A}}$ with $d \geq 3$ and passing to $\mu^1_{\mathcal{A}}$-cohomology in degree zero, one obtains an ordinary $\mathbb{Q}$-linear category, the induced cohomological category $H^0(\mathcal{A})$ – actually, in complete generality $H^0(\mathcal{A})$ may not have identity morphisms, but we will always assume that this is the case (one says that $\mathcal{A}$ is cohomologically unital).

In our application, objects of $\mathcal{A} = \mathcal{F}(M)$ are closed exact Lagrangian submanifolds $L \subset M \setminus \partial M$, with a bit of additional topological structure, namely a grading [14, 22] and a *Spin* structure [9]. If L_0 is transverse to L_1, the space of morphisms $hom_{\mathcal{A}}(L_0, L_1) = CF(L_0, L_1)$ is generated by their intersection points, graded by Maslov index. The composition $\mu^d_{\mathcal{A}}$ counts "pseudo-holomorphic $(d+1)$-gons", which are holomorphic maps from the disk minus $d+1$ boundary points to M. The sides of the "polygons" lie on Lagrangian submanifolds, and the corners are specified intersection points; see Figure 1. There are some technical issues having to do with transversality, which can be solved by a small inhomogeneous perturbation of the holomorphic map equation. This works for all exact symplectic manifolds with contact type boundary, satisfying $c_1 = 0$, and is quite an easy construction by today's standards, since the exactness condition removes the most serious problems (bubbling, obstructions).

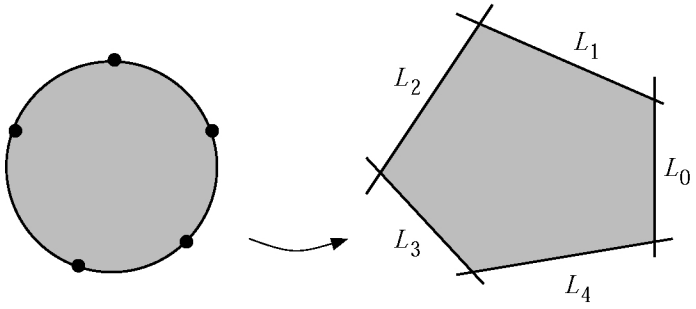


Figure 1:

It is worth while emphasizing that, unlike the case of Gromov-Witten invariants, each one of the coefficients which make up $\mu^d_{\mathcal{A}}$ depends on the choice of perturbation. Only by looking at all of them together does one get an object which is invariant up to a suitable notion of quasi-isomorphism. To get something which is well-defined in a strict sense, one can descend to the cohomological category $H^0(\mathcal{F}(M))$ (which was considered by Donaldson before Fukaya's work) whose morphisms are the Floer cohomology groups, with composition given by the "pair-of-pants" product; but that is rather a waste of information.

At this point, we must admit that there is essentially no chance of computing $\mathcal{F}(M)$ explicitly. The reason is that we know too little about exact Lagrangian submanifolds; indeed, this field contains some of the hardest open questions in symplectic geometry. One way out of this difficulty, proposed by Kontsevich [14], is to make the category more accessible by enlarging it, adding new objects in a formal process, which resembles the introduction of chain complexes over an additive

category. This can be done for any A_∞-category $\mathcal{A}$, and the outcome is called the A_∞-category of twisted complexes, $Tw(\mathcal{A})$. It contains the original A_∞-category as a full subcategory, but this subcategory is not singled out intrinsically, and very different $\mathcal{A}$ can have the same $Tw(\mathcal{A})$. The cohomological category $D^b(\mathcal{A}) = H^0(Tw(\mathcal{A}))$, usually called the derived category of $\mathcal{A}$, is triangulated (passage to cohomology is less damaging at this point, since the triangulated structure allows one to recover many of the higher order products on $Tw(\mathcal{A})$ as Massey products). For our purpose it is convenient to make another enlargement, which is Karoubi or idempotent completion, and leads to a bigger A_∞-category $Tw^\pi(\mathcal{A}) \supset Tw(\mathcal{A})$ and triangulated category $D^\pi(\mathcal{A}) = H^0(Tw^\pi(\mathcal{A}))$. The main property of $D^\pi(\mathcal{A})$ is that for any object X and idempotent endomorphism $\pi : X \to X$, $\pi^2 = \pi$, there is a direct splitting $X = im(\pi) \oplus ker(\pi)$. The details, which are not difficult, will be explained elsewhere.

3. Picard-Lefschetz theory

We will now restrict the class of symplectic manifolds even further:

Assumption 2. *In the situation of Assumption 1, suppose that X is itself a hyperplane section in a smooth projective variety Y, with $\mathcal{K}_Y \cong \mathcal{O}_Y(-X)$. Moreover, $X = X_0$ should be part of a Lefschetz pencil of such sections $\{X_z\}$, whose base locus is $D = X_0 \cap X_\infty$.*

This gives a natural source of Lagrangian spheres in M, namely the vanishing cycles of the Lefschetz pencil. Recall that to any Lagrangian sphere S one can associate a Dehn twist, or Picard-Lefschetz monodromy map, which is a symplectic automorphism τ_S. The symplectic geometry of these maps is quite rich, and contains information which is not visible on the topological level [20, 21, 22]. The action of τ_S on the Fukaya category is encoded in an exact triangle in $Tw(\mathcal{F}(M))$, of the form

$$\begin{array}{ccc} L & \longrightarrow & \tau_S(L) \\ & {\scriptstyle [1]} \searrow \quad \swarrow & \\ & HF^*(S,L) \otimes S & \end{array} \tag{3}$$

for any L, and where the $\otimes$ is just a direct sum of several copies of S in various degrees. This is a consequence of the long exact sequence in Floer cohomology [23].

In the situation of Assumption 2, if we choose a distinguished basis of vanishing cycles $S_1, \dots, S_m$ for the pencil, the product of their Dehn twists is almost the identity map. More precisely, taking into account the "grading" of the objects of the Fukaya category, one finds that

$$\tau_{S_1} \dots \tau_{S_m}(L) \cong L[2]$$

where [2] denotes change in the grading by 2. By combining this trick with (3) one can prove the following result:

Theorem 3. *$S_1, \dots, S_m$ are split-generators for $D^\pi(\mathcal{F}(M))$. This means that any object of $Tw^\pi(\mathcal{F}(M))$ can be obtained from them, up to quasi-isomorphism, by repeatedly forming mapping cones and idempotent splittings.*

4. Hochschild cohomology

The Hochschild cohomology $HH^*(\mathcal{A},\mathcal{A})$ of an A_∞-category $\mathcal{A}$ can be defined by generalizing the Hochschild complex for algebras in a straightforward way, or more elegantly using the A_∞-category $fun(\mathcal{A},\mathcal{A})$ of functors and natural transformations, as endomorphisms of the identity functor. A well-known rather imprecise principle says that "Hochschild cohomology is an invariant of the derived category". In a rigorous formulation which is suitable for our purpose,

$$HH^*(\mathcal{A},\mathcal{A}) \stackrel{?}{\cong} HH^*(Tw^\pi(\mathcal{A}), Tw^\pi(\mathcal{A})). \tag{4}$$

This is unproved at the moment, because $Tw^\pi(\mathcal{A})$ itself has not been considered in the literature before, but it seems highly plausible (a closely related result has been proved in [13]). Hochschild cohomology is important for us because of its role in deformation theory, see the next section; but we want to discuss its possible geometric meaning first.

Let M be as in Assumption 1 (one could more generally take any exact symplectic manifold with contact type boundary and vanishing c_1). Then there is a natural "open-closed string map" from the symplectic cohomology to the Hochschild cohomology of the Fukaya category:

$$SH^*(M) \longrightarrow HH^*(\mathcal{F}(M),\mathcal{F}(M)). \tag{5}$$

This is defined in terms of Riemann surfaces obtained from the disk by removing one interior point and an arbitrary number of boundary points. Near the interior point, one deforms the holomorphic map equation in the same way as in the definition of $SH^*(M)$, using a large Hamiltonian function; otherwise, one uses boundary conditions as for $\mathcal{F}(M)$. Figure 2 shows what the solutions look like.

$HH^*(\mathcal{A},\mathcal{A})$ for any $\mathcal{A}$ carries the structure of a Gerstenhaber algebra, and one can verify that (5) is a morphism of such algebras. Actually, since $SH^*(M)$ is a BV algebra, one expects the same of $HH^*(\mathcal{F}(M),\mathcal{F}(M))$. This should follow from the fact that $\mathcal{F}(M)$ is a cyclic A_∞-category in some appropriate weak sense, but the story has not yet been fully worked out (two relevant papers for the algebraic side are [25] and [24]).

Conjecture 4. *If M is as in Assumption 2, (5) is an isomorphism.*

Assumption 2 appears here mainly for the sake of caution. There are a number of cases which fall outside it, and to which one would want to extend the conjecture, but it is not clear where to draw the line. Certainly, without some restriction on the geometry of M, there can be no connection between the Reeb flow on ∂M and Lagrangian submanifolds?

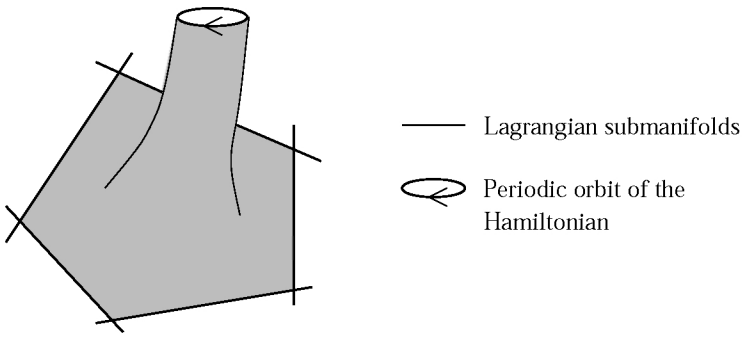


Figure 2:

5. Deformations of categories

The following general definition, due to Kontsevich, satisfies the need for a deformation theory of categories which should be applicable to a wide range of situations: for instance, a deformation of a complex manifold should induce a deformation of the associated differential graded category of complexes of holomorphic vector bundles. By thinking about this example, one quickly realizes that such a notion of deformation must include a change in the set of objects itself. The A_∞-formalism, slightly extended in an entirely natural way, fits that requirement perfectly. The relevance to symplectic topology is less immediately obvious, but it plays a central role in Fukaya, Oh, Ohta and Ono's work on "obstructions" in Floer cohomology [9] (a good expository account from their point of view is [5]).

For concreteness we consider only A_∞-deformations with one formal parameter, that is to say over $\mathbb{Q}[[t]]$. Such a deformation $\mathcal{E}$ is given by a set $Ob\,\mathcal{E}$ of objects, and for any two objects a space $hom_{\mathcal{E}}(X_0, X_1)$ of morphisms which is a free graded $\mathbb{Q}[[t]]$-module, together with composition operations as before but now including a 0-ary one: this consists of a so-called "obstruction cocycle"

$$\mu^0_{\mathcal{E}} \in hom^2_{\mathcal{E}}(X, X) \tag{6}$$

for every object X, and it must be of order t (no constant term). There is a sequence of associativity equations, extending those of an A_∞-category by terms involving $\mu^0_{\mathcal{E}}$. Clearly, if one sets $t = 0$ (by tensoring with $\mathbb{Q}$ over $\mathbb{Q}[[t]]$), $\mu^0_{\mathcal{E}}$ vanishes and the outcome is an ordinary A_∞-category over $\mathbb{Q}$. This is called the special fibre and denoted by $\mathcal{E}_{sp}$. One says that $\mathcal{E}$ is a deformation of $\mathcal{E}_{sp}$.

A slightly more involved construction associates to $\mathcal{E}$ two other A_∞-categories, the global section category $\mathcal{E}_{gl}$ and the generic fibre $\mathcal{E}_{gen}$, which are defined over $\mathbb{Q}[[t]]$ and over the Laurent series ring $\mathbb{Q}[t^{-1}][[t]]$, respectively. One first enlarges $\mathcal{E}$ to a bigger A_∞-deformation $\mathcal{E}_c$ by coupling the existing objects with formal connections (the terminology comes from the application to complexes of vector bundles). Objects of $\mathcal{E}_c$ are pairs (X, α) consisting of $X \in Ob\,\mathcal{E}$ and an $\alpha \in hom^1_{\mathcal{E}}(X, X)$

which must be of order t. The morphism spaces remain the same as in $\mathcal{E}$, but all the composition maps are deformed by infinitely many contributions from the connection. For instance,

$$\mu^0_{\mathcal{E}_c} = \mu^0_{\mathcal{E}} + \mu^1_{\mathcal{E}}(\alpha) + \mu^2_{\mathcal{E}}(\alpha,\alpha) + \cdots \in hom^2_{\mathcal{E}_c}((X,\alpha),(X,\alpha)) = hom^2_{\mathcal{E}}(X,X). \quad (7)$$

$\mathcal{E}_{gl} \subset \mathcal{E}_c$ is the full A_∞-subcategory of objects for which (7) is zero; and $\mathcal{E}_{gen}$ is obtained from this by inverting t. The transition from $\mathcal{E}_{sp}$ to $\mathcal{E}_{gl}$ and $\mathcal{E}_{gen}$ affects the set of objects in the following way: if for some X one cannot find an α such that (7) vanishes, then the object is "obstructed" and does not survive into $\mathcal{E}_{gl}$; if on the other hand there are many different α, a single X can give rise to a whole family of objects of $\mathcal{E}_{gl}$. Finally, two objects of $\mathcal{E}_{gen}$ can be isomorphic even though the underlying objects of $\mathcal{E}_{sp}$ aren't; this happens when the isomorphism involves negative powers of t.

The classification of A_∞-deformations of an A_∞-category $\mathcal{A}$ is governed by its Hochschild cohomology, or rather by the dg Lie algebra underlying $HH^{*+1}(\mathcal{A},\mathcal{A})$, in the sense of general deformation theory [11]. We cannot summarize that theory here, but as a simple example, suppose that $HH^2(\mathcal{A},\mathcal{A}) \cong \mathbb{Q}$. Then a nontrivial A_∞-deformation of $\mathcal{A}$, if it exists, is unique up to equivalence and change of parameter $t \mapsto f(t)$ (to be accurate, $f(t)$ may contain roots of t, so the statement holds over $\mathbb{Q}[[t, t^{1/2}, t^{1/3}, \dots]]$). The intuitive picture is that the "versal deformation space" has dimension ≤ 1, so that any two non-constant arcs in it must agree up to reparametrization.

In the situation of Assumption 1, the embedding of our exact symplectic manifold M into X should give rise to an A_∞-deformation $\mathcal{F}(M \subset X)$. We say "should" because the details, which in general require the techniques of [9], have not been carried out yet. Roughly speaking one takes the same objects as in $\mathcal{F}(M)$ and the same morphism spaces, tensored with $\mathbb{Q}[[t]]$, but now one allows "holomorphic polygons" which map to X, hence may intersect the divisor D. The numbers of such polygons intersecting D with multiplicity k will form the t^k term of the composition maps in $\mathcal{F}(M \subset X)$. Because there can be holomorphic discs bounding our Lagrangian submanifolds in X, nontrivial obstruction cocycles (6) may appear.

The intended role of $\mathcal{F}(M \subset X)$ is to interpolate between $\mathcal{F}(M)$, which we have been mostly discussing up to now, and the Fukaya category $\mathcal{F}(X)$ of the closed symplectic manifold X as defined in [9, 6]. The t^0 coefficients count polygons which are disjoint from D, and these will automatically lie in M, so that

$$\mathcal{F}(M \subset X)_{sp} \cong \mathcal{F}(M).$$

The relation between the generic fibre and $\mathcal{F}(X)$ is less straightforward. First of all, $\mathcal{F}(M \subset X)_{gen}$ will be an A_∞-category over $\mathbb{Q}[t^{-1}, t]]$, whereas $\mathcal{F}(X)$ is defined over the Novikov ring Λ_t. Intuitively, one can think of this difference as the consequence of a singular deformation of the symplectic form. Namely, if one takes a sequence of symplectic forms (all in the same cohomology class) converging towards the current $[D]$, the symplectic areas of holomorphic discs u would tend to the intersection number $u \cdot D$. A more serious issue is that $\mathcal{F}(M \subset X)_{gen}$ is clearly smaller than

$\mathcal{F}(X)$, because it contains only Lagrangian submanifolds which lie in M. However, that difference may disappear if one passes to derived categories:

Conjecture 5. *In the situation of Assumption 2, there is a canonical equivalence of triangulated categories*

$$D^\pi(\mathcal{F}(M \subset X)_{gen} \otimes_{\mathbb{Q}[t^{-1}][[t]]} \Lambda_t) \cong D^\pi(\mathcal{F}(X)).$$

In comparison with the previous conjecture, Assumption 2 is far more important here. The idea is that there should be an analogue of Theorem 3 for $D^\pi(\mathcal{F}(X))$, saying that this category is split-generated by vanishing cycles, hence by objects which are also present in $\mathcal{F}(M \subset X)$.

To pull together the various speculations, suppose that $Y = \mathbb{CP}^{n+1}$ for some $n \geq 3$; $X \subset Y$ is a hypersurface of degree $n+2$; and $D \subset X$ is the intersection of two such hypersurfaces. Then $D^\pi(\mathcal{F}(M))$ is split-generated by finitely many objects, hence $Tw^\pi(\mathcal{F}(M))$ is at least in principle accessible to computation. Conjecture 4 together with (2), (4) tells us that $HH^2(\mathcal{F}(M), \mathcal{F}(M)) \cong HH^2(Tw^\pi(\mathcal{F}(M)), Tw^\pi(\mathcal{F}(M)))$ is at most one-dimensional, so an A_∞-deformation of $Tw^\pi(\mathcal{F}(M))$ is unique up to a change of the parameter t. From this deformation, Conjecture 5 would enable one to find $D^\pi(\mathcal{F}(X))$, again with the indeterminacy in the parameter (fixing this is somewhat like computing the mirror map).

Acknowledgements. Obviously, the ideas outlined here owe greatly to Fukaya and Kontsevich. The author is equally indebted to Auroux, Donaldson, Getzler, Joyce, Khovanov, Smith, and Thomas (an incomplete list), all of whom have influenced his thinking considerably. The preparation of this talk at the Institute for Advanced Study was supported by NSF grant DMS-9729992.

References

[1] M. Chas and D. Sullivan, *String topology*, Preprint math.GT/9911159.

[2] K. Cieliebak, A. Floer, and H. Hofer, *Symplectic homology II: a general construction*, Math. Z. **218** (1995), 103–122.

[3] Ya. Eliashberg, A. Givental, and H. Hofer, *Introduction to symplectic field theory*, Geom. Funct. Anal. **Special Volume, Part II** (2000), 560–673.

[4] K. Fukaya, *Asymptotic analysis, multivalued Morse theory, and mirror symmetry*, Preprint 2002.

[5] ———, *Deformation theory, homological algebra, and mirror symmetry*, Preprint, December 2001.

[6] ———, *Floer homology and mirror symmetry II*, Preprint 2001.

[7] ———, *Morse homotopy, A_∞-categories, and Floer homologies*, Proceedings of GARC workshop on Geometry and Topology (H. J. Kim, ed.), Seoul National University, 1993.

[8] K. Fukaya and Y.-G. Oh, *Zero-loop open strings in the cotangent bundle and Morse homotopy*, Asian J. Math. **1** (1998), 96–180.

[9] K. Fukaya, Y.-G. Oh, H. Ohta, and K. Ono, *Lagrangian intersection Floer theory - anomaly and obstruction*, Preprint, 2000.

[10] E. Getzler, *Batalin-Vilkovisky algebras and 2d Topological Field Theories*, Commun. Math. Phys **159** (1994), 265–285.
[11] W. Goldman and J. Millson, *The deformation theory of the fundamental group of compact Kähler manifolds*, IHES Publ. Math. **67**, 43–96.
[12] E.-N. Ionel and T. Parker, *Gromov-Witten invariants of symplectic sums*, Math. Res. Lett. **5** (1998), 563–576.
[13] B. Keller, *Invariance and localization for cyclic homology of DG algebras*, J. Pure Appl. Alg. **123** (1998), 223–273.
[14] M. Kontsevich, *Homological algebra of mirror symmetry*, Proceedings of the International Congress of Mathematicians (Zürich, 1994), Birkhäuser, 1995, 120–139.
[15] M. Kontsevich and Y. Soibelman, *Homological mirror symmetry and torus fibrations*, Symplectic geometry and mirror symmetry, World Scientific, 2001, 203–263.
[16] A.-M. Li and Y. Ruan, *Symplectic surgery and Gromov-Witten invariants of Calabi-Yau 3-folds*, Invent. Math. **145** (2001), 151–218.
[17] P. Ozsvath and Z. Szabo, *Holomorphic disks and topological invariants for rational homology three-spheres*, Preprint math.SG/0101206.
[18] A. Polishchuk and E. Zaslow, *Categorical mirror symmetry: the elliptic curve*, Adv. Theor. Math. Phys. **2** (1998), 443–470.
[19] M. Poźniak, *Floer homology, Novikov rings and clean intersections*, Northern California Symplectic Geometry Seminar, Amer. Math. Soc., 1999, 119–181.
[20] P. Seidel, *Floer homology and the symplectic isotopy problem*, Ph.D. thesis, Oxford University, 1997.
[21] ———, *Lagrangian two-spheres can be symplectically knotted*, J. Differential Geom. **52** (1999), 145–171.
[22] ———, *Graded Lagrangian submanifolds*, Bull. Soc. Math. France **128** (2000), 103–146.
[23] ———, *A long exact sequence for symplectic Floer cohomology*, Preprint math.SG/0105186.
[24] D. Tamarkin and B. Tsygan, *Noncommutative differential calculus, homotopy BV algebras and formality conjectures*, Preprint math.KT/0002116.
[25] T. Tradler, *Infinity-inner-products on A-infinity-algebras*, Preprint math.-AT/0108027.
[26] C. Viterbo, *Functors and computations in Floer homology with applications, Part I*, Geom. Funct. Anal. **9** (1999), 985–1033.
[27] ———, *Functors and computations in Floer homology with applications, Part II*, Preprint 1996.

ICM 2002 · Vol. II · 361–369

Heat Kernels and the Index Theorems on Even and Odd Dimensional Manifolds*

Weiping Zhang†

Abstract

In this talk, we review the heat kernel approach to the Atiyah-Singer index theorem for Dirac operators on closed manifolds, as well as the Atiyah-Patodi-Singer index theorem for Dirac operators on manifolds with boundary. We also discuss the odd dimensional counterparts of the above results. In particular, we describe a joint result with Xianzhe Dai on an index theorem for Toeplitz operators on odd dimensional manifolds with boundary.

2000 Mathematics Subject Classification: 58G.
Keywords and Phrases: Index theorems, heat kernels, eta-invariants, Toeplitz operators.

1. Introduction

As is well-known, the index theorem proved by Atiyah and Singer [AS1] in 1963, which expresses the analytically defined index of elliptic differential operators through purely topological terms, has had a wide range of implications in mathematics as well as in mathematical physics. Moreover, there have been up to now many different proofs of this celebrated result.

The existing proofs of the Atiyah-Singer index theorem can roughly be divided into three categories:

(i) The cobordism proof: this is the proof originally given in [AS1]. It uses the cobordism theory developed by Thom and modifies Hirzebruch's proof of his Signature theorem as well as his Riemann-Roch theorem;

(ii) The K-theoretic proof: this is the proof given by Atiyah and Singer in [AS2]. It modifies Grothendieck's proof of the Hirzebruch-Riemann-Roch theorem and relies on the topological K-theory developed by Atiyah and Hirzebruch. The Bott periodicity theorem plays an important role in this proof;

*Partially supported by the MOEC and the 973 Project.
†Nankai Institute of Mathematics, Nankai University, Tianjin 300071, China. E-mail: weiping@nankai.edu.cn

(iii) The heat kernel proof: this proof originates from a simple and beautiful formula due to Mckean and Singer [MS], and has closer relations with differential geometry as well as mathematical physics. It also lead directly to the important Atiyah-Patodi-Singer index theorem for Dirac operators on manifolds with boundary.

In this article, we will survey some of the developments concerning the heat kernel proofs of various index theorems, including a recent result with Dai [DZ2] on an index theorem for Toeplitz operators on odd dimensional manifolds with boundary.

2. Heat kernels and the index theorems on even dimensional manifolds

We start with a smooth closed oriented $2n$-dimensional manifold M and two smooth complex vector bundles E, F over M, on which there is an elliptic differential operator between the spaces of smooth sections, $D_+ : \Gamma(E) \to \Gamma(F)$.

If we equip TM with a Riemannian metric and E, F with Hermitian metrics repectively, then $\Gamma(E)$ and $\Gamma(F)$ will carry canonically induced inner products.

Let $D_- : \Gamma(F) \to \Gamma(E)$ be the formal adjoint of D_+ with respect to these inner products. Then the index of D_+ is given by

$$\operatorname{ind} D_+ = \dim(\ker D_+) - \dim(\ker D_-). \tag{2.1}$$

It is a topological invariant not depending on the metrics on TM, E and F.

The famous Mckean-Singer formula [MS] says that $\operatorname{ind} D_+$ can also be computed by using the heat operators associated to the Laplacians D_-D_+ and D_+D_-. That is, for any $t > 0$, one has

$$\operatorname{ind} D_+ = \operatorname{Tr}[\exp(-tD_-D_+)] - \operatorname{Tr}[\exp(-tD_+D_-)]. \tag{2.2}$$

By introducing the $\mathbf{Z}_2$-graded vector bundle $E \oplus F$ and setting $D = \begin{pmatrix} 0 & D_- \\ D_+ & 0 \end{pmatrix}$, we can rewrite the difference of the two traces in the right hand side of (2.2) as a single "supertrace" as follows,

$$\operatorname{ind} D_+ = \operatorname{Tr}_s\left[\exp\left(-tD^2\right)\right], \quad \text{for any } t > 0. \tag{2.2}'$$

Let $P_t(x, y)$ be the smooth kernel of $\exp(-tD^2)$ with respect to the volume form on M. For any $f \in \Gamma(E \oplus F)$, one has

$$\exp\left(-tD^2\right) f(x) = \int_M P_t(x, y) f(y) dy. \tag{2.3}$$

In particular,

$$\operatorname{Tr}_s\left[\exp\left(-tD^2\right)\right] = \int_M \operatorname{Tr}_s[P_t(x, x)]\, dx. \tag{2.4}$$

Now, for simplicity, we assume that the elliptic operator D is of order one. Then by a standard result, which goes back to Minakshisundaram and Pleijiel [MP], one has that when $t > 0$ tends to 0,

$$P_t(x,x) = \frac{1}{(4\pi t)^n}\left(a_{-n} + a_{-n+1}t + \cdots + a_0 t^n + o_x\left(t^n\right)\right), \tag{2.5}$$

where $a_i \in \mathrm{End}((E \oplus F)_x)$, $i = -n, \ldots, 0$.

By (2.2)', (2.4) and (2.5), and by taking $t > 0$ small enough, one deduces that

$$\int_M \mathrm{Tr}_s[a_i]dx = 0, \quad -n \leq i < 0,$$

$$\mathrm{ind}\, D_+ = \left(\frac{1}{4\pi}\right)^n \int_M \mathrm{Tr}_s[a_0]dx. \tag{2.6}$$

Mckean and Singer conjectured in [MS] that for certain geometric operators, there should be some "fantastic cancellation" so that the following far reaching refinement of (2.6) holds,

$$\mathrm{Tr}_s[a_i] = 0, \quad -n \leq i < 0,$$

and moreover, $\mathrm{Tr}_s[a_0]$ can be calculated simply in the Chern-Weil geometric theory of characteristic classes.

In fact, as a typical example, let M be an even dimensional compact smooth oriented *spin* manifold carrying a Riemannian metric g^{TM}. Let R^{TM} be the curvature of the Levi-Civita connection associated to g^{TM}. Let $S(TM) = S_+(TM) \oplus S_-(TM)$ be the Hermitian vector bundle of (TM, g^{TM})-spinors, and $D_+ : \Gamma(S_+(TM)) \to \Gamma(S_-(TM))$ the associated *Dirac* operator.

One then has the formula (cf. [BGV, Chap. 4, 5]),

$$\lim_{t\to 0} \mathrm{Tr}_s\left[P_t(x,x)\right]dx = \left\{\widehat{A}\left(\frac{R^{TM}}{2\pi}\right)\right\}^{\max} := \left\{\det{}^{1/2}\left(\frac{\frac{\sqrt{-1}}{4\pi}R^{TM}}{\sinh\left(\frac{\sqrt{-1}}{4\pi}R^{TM}\right)}\right)\right\}^{\max}, \tag{2.7}$$

which implies the Atiyah-Singer index theorem [AS1] for D_+:

$$\mathrm{ind}\, D_+ = \widehat{A}(M) := \int_M \widehat{A}\left(\frac{R^{TM}}{2\pi}\right). \tag{2.8}$$

A result of type (2.7) is called a *local index theorem.* The first proof of such a local result was given by V. K. Patodi [P] for the de Rham-Hodge operator $d + d^*$. Other direct heat kernel proofs of (2.7) have been given by Berline-Vergne, Bismut, Getzler and Yu respectively. We refer to [BGV] and [Yu] for more details.

The heat kernel proof of the local index theorem leads to a generalization of the index theorem for Dirac operators to the case of manifolds with boundary. This was achieved by Atiyah, Patodi and Singer in [APS], and will be reviewed in the next section.

3. The index theorem for Dirac operators on even dimensional manifolds with boundary

Let M be a smooth compact oriented even dimensional *spin* manifold with (nonempty) smooth boundary ∂M. Then ∂M is again oriented and *spin*.

Let g^{TM} be a metric on TM. Let $g^{T\partial M}$ be its restriction on $T\partial M$. We assume for simplicity that g^{TM} is of *product structure* near the boundary ∂M. Let $S(TX) = S_+(TX) \oplus S_-(TX)$ be the $\mathbf{Z}_2$-graded Hermitian vector bundle of (TX, g^{TX})-spinors.

Since now M has a nonempty boundary ∂M, the associated Dirac operator $D_+ : \Gamma(S_+(TM)) \to \Gamma(S_-(TM))$ is *not* elliptic. To get an elliptic problem, one needs to introduce an elliptic boundary condition for D_+, and this was achieved by Atiyah, Patodi and Singer in [APS]. It is remarkable that this boundary condition, to be described right now, is *global* in nature.

First of all, the Dirac operator D_+ induces canonically a formally self-adjoint first order elliptic differential operator

$$D_{\partial M} : \Gamma\left(S_+(TM)|_{\partial M}\right) \to \Gamma\left(S_+(TM)|_{\partial M}\right),$$

which is called the induced Dirac operator on the boundary ∂M.

Clearly, the L^2-completion of $S_+(TM)|_{\partial M}$ admits an orthogonal decomposition

$$L^2\left(S_+(TM)|_{\partial X}\right) = \bigoplus_{\lambda \in \mathrm{Spec}(D_{\partial M})} E_\lambda, \tag{3.1}$$

where E_λ is the eigenspace of λ.

Let $L^2_{\geq 0}(S_+(TM)|_{\partial M})$ denote the direct sum of the eigenspaces E_λ associated to the eigenvalues $\lambda \geq 0$. Let $P_{\geq 0}$ denote the orthogonal projection from $L^2(S_+(TM)|_{\partial M})$ to $L^2_{\geq 0}(S_+(TM)|_{\partial M})$. We call $P_{\geq 0}$ the *Atiyah-Patodi-Singer projection* associated to $D_{\partial M}$, to emphasize its role in [APS].

Then by [APS], the boundary problem

$$(D_+, P_{\geq 0}) : \{u : u \in \Gamma(S_+(TM)),\ P_{\geq 0}\left(u|_{\partial M}\right) = 0\} \to \Gamma(S_-(TM)), \tag{3.2}$$

is Fredholm. We call this elliptic boundary problem the Atiyah-Patodi-Singer boundary problem associated to D_+. We denote by $\mathrm{ind}\,(D_+, P_{\geq 0})$ the index of the Fredholm operator (3.2).

The Atiyah-Patodi-Singer index theorem *The following identity holds,*

$$\mathrm{ind}\,(D_+, P_{\geq 0}) = \int_M \widehat{A}\left(\frac{R^{TM}}{2\pi}\right) - \overline{\eta}\left(D_{\partial M}\right). \tag{3.3}$$

The boundary correction term $\overline{\eta}(D_{\partial M})$ appearing in the right hand side of (3.3) is a spectral invariant associated to the induced Dirac operator $D_{\partial M}$ on ∂M. It is defined as follows: for any complex number $s \in \mathbf{C}$ with $\mathrm{Re}(s) > \dim M$, define

$$\eta\left(D_{\partial M}, s\right) = \sum_{\lambda \in \mathrm{Spec}(D_{\partial M})} \frac{\mathrm{sgn}(\lambda)}{|\lambda|^s}. \tag{3.4}$$

By using the heat kernel method, one can show easily that $\eta(D_{\partial M}, s)$ can be extended to a meromorphic function on $\mathbf{C}$, which is holomorphic at $s = 0$. Following [APS], we then define

$$\overline{\eta}\,(D_{\partial M}) = \frac{\dim\,(\ker D_{\partial M}) + \eta\,(D_{\partial M}, 0)}{2} \tag{3.5}$$

and call it the (reduced) eta invariant of $D_{\partial M}$.

The eta invariants of Dirac operators have played important roles in many aspects of topology, geometry and mathematical physics.

In the next sections, we will discuss the role of eta invariants in the heat kernel approaches to the index theorems on odd dimensional manifolds.

4. Heat kernels and the index theorem on odd dimensional manifolds

Let M be now an *odd* dimensional smooth closed oriented *spin* manifold. Let g^{TM} be a Riemannian metric on TM and $S(TM)$ the associated Hermitian vector bundle of (TM, g^{TM})-spinors.[1] In this case, the associated *Dirac* operator $D : \Gamma(TM) \to \Gamma(TM)$ is (formally) *self-adjoint.*[2] Thus, one can proceed as in Section 3 to construct the Atiyah-Patodi-Singer projection

$$P_{\geq 0} : L^2(S(TM)) \to L^2_{\geq 0}(S(TM)).$$

Now consider the trivial vector bundle $\mathbf{C}^N$ over M. We equip $\mathbf{C}^N$ with the canonical trivial metric and connection. Then $P_{\geq 0}$ extends naturally to an orthogonal projection from $L^2(S(TM) \otimes \mathbf{C}^N)$ to $L^2_{\geq 0}(S(TM) \otimes \mathbf{C}^N)$ by acting as identity on $\mathbf{C}^N$. We still denote this extension by $P_{\geq 0}$.

On the other hand, let

$$g : M \to U(N)$$

be a smooth map from M to the unitary group $U(N)$. Then g can be interpreted as automorphism of the trivial complex vector bundle $\mathbf{C}^N$. Moreover g extends naturally to an action on $L^2(S(TM) \otimes \mathbf{C}^N)$ by acting as identity on $L^2(S(TM))$. We still denote this extended action by g.

With the above data given, one can define a *Toeplitz* operator T_g as follows,

$$T_g = P_{\geq 0} g P_{\geq 0} : L^2_{\geq 0}\left(S(TM) \otimes \mathbf{C}^N\right) \longrightarrow L^2_{\geq 0}\left(S(TM) \otimes \mathbf{C}^N\right). \tag{4.1}$$

The first important fact is that T_g is a Fredholm operator. Moreover, it is equivalent to an elliptic pseudodifferential operator of order zero. Thus one can compute its index by using the Atiyah-Singer index theorem [AS2], as was indicated in the paper of Baum and Douglas [BD], and the result is

$$\operatorname{ind} T_g = -\left\langle \widehat{A}(TM)\mathrm{ch}(g), [M] \right\rangle, \tag{4.2}$$

[1] Since now M is of odd dimension, the bundle of spinors does not admit a $\mathbf{Z}_2$-graded structure.

[2] In fact, if M bounds an even dimensional spin manifold, then D can be thought of as the induced Dirac operator on boundary appearing in the previous section.

where $\mathrm{ch}(g)$ is the odd Chern character associated to g.

There is also an analytic proof of (4.2) by using heat kernels. For this one first applies a result of Booss and Wojciechowski (cf. [BW]) to show that the computation of $\mathrm{ind}\, T_g$ is equivalent to the computation of the spectral flow of the linear family of self-adjoint elliptic operators, acting of $\Gamma(S(TM) \otimes \mathbf{C}^N)$, which connects D and gDg^{-1}. The resulting spectral flow can then be computed by variations of η-invariants, where the heat kernels are naturally involved.

The above ideas have been extended in [DZ1] to give a heat kernel proof of a family extension of (4.2).

5. An index theorem for Toeplitz operators on odd dimensional manifolds with boundary

In this section, we describe an extension of (4.2) to the case of manifolds with boundary, which was proved recently in my paper with Xianzhe Dai [DZ2]. This result can be thought of as an odd dimensional analogue of the Atiyah-Patodi-Singer index theorem described in Section 3.

This section is divided into three subsections. In Subsection 4.1, we extend the definition of Toeplitz operators to the case of manifolds with boundary. In Subsection 4.2, we define an η-invariant for cylinders which will appear in the statement of the main result to be described in Subsection 4.3.

5.1. Toeplitz operators on manifolds with boundary

Let M be an odd dimensional oriented *spin* manifold with (nonempty) boundary ∂M. Then ∂M is also oriented and spin. Let g^{TM} be a Riemannian metric on TM such that it is of product structure near the boundary ∂M. Let $S(TM)$ be the Hermitian bundle of spinors associated to (M, g^{TM}). Since $\partial M \neq \emptyset$, the *Dirac* operator $D : \Gamma(S(TM)) \to \Gamma(S(TM))$ is no longer elliptic. To get an elliptic operator, one needs to impose suitable boundary conditions, and it turns out that again we will adopt the boundary conditions introduced by Atiyah, Patodi and Singer [APS].

Let $D_{\partial M} : \Gamma(S(TM)|_{\partial M}) \to \Gamma(S(TM)|_{\partial M})$ be the canonically induced Dirac operator on the boundary ∂M. Then $D_{\partial M}$ is elliptic and (formally) self-adjoint. For simplicity, we assume here that $D_{\partial M}$ is *invertible*, that is, $\ker D_{\partial M} = 0$.

Let $P_{\partial M, \geq 0}$ denote the Atiyah-Patodi-Singer projection from $L^2(S(TM)|_{\partial M})$ to $L^2_{\geq 0}(S(TM)|_{\partial M})$. Then $(D, P_{\partial M, \geq 0})$ forms a *self-adjoint* elliptic boundary problem. We will also denote the corresponding elliptic self-adjoint operator by $D_{P_{\partial M, \geq 0}}$.

Let $L^2_{P_{\partial M, \geq 0}, \geq 0}(S(TM))$ be the space of the direct sum of eigenspaces of non-negative eigenvalues of $D_{P_{\partial M, \geq 0}}$. Let $P_{P_{\partial M, \geq 0}, \geq 0}$ denote the orthogonal projection from $L^2(S(TM))$ to $L^2_{P_{\partial M, \geq 0}, \geq 0}(S(TM))$.

Now let $\mathbf{C}^N$ be the trivial complex vector bundle over M of rank N, which carries the trivial Hermitian metric and the trivial Hermitian connection. We extend $P_{P_{\partial M, \geq 0}, \geq 0}$ to act as identity on $\mathbf{C}^N$.

Let $g : M \to U(N)$ be a smooth unitary automorphism of $\mathbf{C}^N$. Then g extends to an action on $S(TM) \otimes \mathbf{C}^N$ by acting as identity on $S(TM)$.

Since g is unitary, one verifies easily that the operator $gP_{\partial M,\geq 0}g^{-1}$ is an orthogonal projection on $L^2((S(TM)\otimes\mathbf{C}^N)|_{\partial M})$, and that $gP_{\partial M,\geq 0}g^{-1}-P_{\partial M,\geq 0}$ is a pseudodifferential operator of order less than zero. Moreover, the pair $(D, gP_{\partial M,\geq 0}g^{-1})$ forms a *self-adjoint* elliptic boundary problem. We denote its associated elliptic self-adjoint operator by $D_{gP_{\partial M,\geq 0}g^{-1}}$.

Let $L^2_{gP_{\partial M,\geq 0}g^{-1},\geq 0}(S(TM)\otimes\mathbf{C}^N)$ be the space of the direct sum of eigenspaces of nonnegative eigenvalues of $D_{gP_{\partial M,\geq 0}g^{-1}}$. Let $P_{gP_{\partial M,\geq 0}g^{-1},\geq 0}$ denote the orthogonal projection from $L^2(S(TM)\otimes\mathbf{C}^N)$ to $L^2_{gP_{\partial M,\geq 0}g^{-1},\geq 0}(S(TM)\otimes\mathbf{C}^N)$.

Clearly, if $s\in L^2(S(TM)\otimes\mathbf{C}^N)$ verifies $P_{\partial M,\geq 0}(s|_{\partial M})=0$, then gs verifies

$$gP_{\partial M,\geq 0}g^{-1}\left((gs)|_{\partial M}\right)=0.$$

Definition 5.1 The *Toeplitz operator* T_g is defined by

$$T_g = P_{gP_{\partial M,\geq 0}g^{-1},\geq 0}gP_{P_{\partial M,\geq 0},\geq 0}:$$

$$L^2_{P_{\partial M,\geq 0},\geq 0}\left(S(TM)\otimes\mathbf{C}^N\right)\to L^2_{gP_{\partial M,\geq 0}g^{-1},\geq 0}\left(S(TM)\otimes\mathbf{C}^N\right).$$

One verifies that T_g is a Fredholm operator. The main result of this section evaluates the index of T_g by more geometric quantities.

5.2. An η-invariant associated to g

We consider the cylinder $[0,1]\times\partial M$. Clearly, the restriction of g on ∂M extends canonically to this cylinder.

Let $D|_{[0,1]\times\partial M}$ be the restriction of D on $[0,1]\times\partial M$. We equip the boundary condition $P_{\partial M,\geq 0}$ at $\{0\}\times\partial M$ and the boundary condition $\mathrm{Id}-gP_{\partial M,\geq 0}g^{-1}$ at $\{1\}\times\partial M$. Then $(D|_{[0,1]\times\partial M}, P_{\partial M,\geq 0}, \mathrm{Id}-gP_{\partial M,\geq 0}g^{-1})$ forms a self-adjoint elliptic boundary problem. We denote the corresponding elliptic self-adjoint operator by $D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}}$.

Let $\eta(D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}},s)$ be the η-function of $s\in\mathbf{C}$ which, when $\mathrm{Re}(s)>>0$, is defined by

$$\eta\left(D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}},s\right)=\sum_{\lambda\neq 0}\frac{\mathrm{sgn}(\lambda)}{|\lambda|^s},$$

where λ runs through the nonzero eigenvalues of $D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}}$.

It is proved in [DZ2] that under our situation, $\eta(D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}},s)$ can be extended to a meromorphic function on $\mathbf{C}$ which is holomorphic at $s=0$.

Let $\overline{\eta}(D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}})$ be the reduced η-invariant defined by

$$\overline{\eta}\left(D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}}\right)$$

$$=\frac{\dim\ker\left(D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}}\right)+\eta\left(D_{P_{\partial M,\geq 0},gP_{\partial M,\geq 0}g^{-1}}\right)}{2}.$$

5.3. An index theorem for T_g

Let ∇^{TM} be the Levi-Civita connection associated to the Riemannian metric g^{TM}. Let $R^{TM} = (\nabla^{TM})^2$ be the curvature of ∇^{TM}. Also, we use d to denote the trivial connection on the trivial vector bundle $\mathbf{C}^N$ over M. Then $g^{-1}dg$ is a $\Gamma(\mathrm{End}(\mathbf{C}^N))$ valued 1-form over M.

Let $\mathrm{ch}(g,d)$ denote the odd Chern character form (cf. [Z]) of (g,d) defined by

$$\mathrm{ch}(g,d) = \sum_{n=0}^{(\dim M-1)/2} \frac{n!}{(2n+1)!}\left(\frac{1}{2\pi\sqrt{-1}}\right)^{n+1} \mathrm{Tr}\left[\left(g^{-1}dg\right)^{2n+1}\right].$$

Let $\mathcal{P}_M$ denote the Calderón projection associated to D on M (cf. [BW]). Then $\mathcal{P}_M$ is an orthogonal projection on $L^2((S(TM)\otimes\mathbf{C}^N)|_{\partial M})$, and that $\mathcal{P}_M - P_{\partial M,\geq 0}$ is a pseudodifferential operator of order less than zero.

Let $\tau_\mu(P_{\partial M,\geq 0}, gP_{\partial M,\geq 0}g^{-1}, \mathcal{P}_M) \in \mathbf{Z}$ be the Maslov triple index in the sense of Kirk and Lesch [KL, Definition 6.8].

We can now state the main result of [DZ2], which generalizes an old result of Douglas and Wojciechowski [DoW], as follows.

Theorem 5.2 *The following identity holds,*

$$\begin{aligned}\mathrm{ind}\, T_g = &-\int_M \widehat{A}\left(\frac{R^{TM}}{2\pi}\right)\mathrm{ch}(g,d) + \overline{\eta}\left(D_{P_{\partial M,\geq 0}, gP_{\partial M,\geq 0}g^{-1}}\right)\\ &-\tau_\mu\left(P_{\partial M,\geq 0}, gP_{\partial M,\geq 0}g^{-1}, \mathcal{P}_M\right).\end{aligned}$$

The following immediate consequence is of independent interests.

Corollary 5.3 *The number*

$$\int_M \widehat{A}\left(\frac{R^{TM}}{2\pi}\right)\mathrm{ch}(g,d) - \overline{\eta}\left(D_{P_{\partial M,\geq 0}, gP_{\partial M,\geq 0}g^{-1}}\right)$$

is an integer.

The strategy of the proof of Theorem 5.2 given in [DZ2] is the same as that of the heat kernel proof of (4.2). However, due to the appearance of the boundary ∂M, one encounters new difficulties. To overcome these difficulties, one makes use of the recent result on the splittings of η invariants (cf. [KL]) as well as some ideas involved in the Connes-Moscovici local index theorem in noncommutative geometry [CM] (see also [CH]). Moreover, the local index calculations appearing near ∂M is highly nontrivial. We refer to [DZ2] for more details.

References

[APS] M. F. Atiyah, V. K. Patodi and I. M. Singer, Spectral asymmetry and Riemannian geometry I. *Proc. Cambridge Philos. Soc.* 77 (1975), 43–69.

[AS1] M. F. Atiyah and I. M. Singer, The index of elliptic operators on compact manifolds. *Bull. Amer. Math. Soc.* 69 (1963), 422–433.

[AS2] M. F. Atiyah and I. M. Singer, The index of elliptic operators I. *Ann. of Math.* 87 (1968), 484–530.

[BD] P. Baum and R. G. Douglas, K-homology and index theory, in *Proc. Sympos. Pure and Appl. Math.*, Vol. 38, 117–173, Amer. Math. Soc. Providence, 1982.

[BGV] N. Berline, E. Getzler and M. Vergne, *Heat Kernels and Dirac operators.* Grundlagen der Math. Wissenschften Vol. 298. Springer-Verlag, 1991.

[BW] B. Booss and K. Wojciechowski, *Elliptic Boundary Problems for Dirac Operators*, Birkhäuser, 1993.

[CH] S. Chern and X. Hu, Equivariant Chern character for the invariant Dirac operator. *Michigan Math. J.* 44 (1997), 451–473.

[CM] A. Connes and H. Moscovici, The local index formula in noncommutative geometry. *Geom. Funct. Anal.* 5 (1995), 174–243.

[DZ1] X. Dai and W. Zhang, Higher spectral flow. *J. Funct. Anal.* 157 (1998), 432–469.

[DZ2] X. Dai and W. Zhang, An index theorem for Toeplitz operators on odd dimensional manifolds with boundary. *Preprint*, math.DG/0103230.

[DoW] R. G. Douglas and K. P. Wojciechowski, Adiabatic limits of the η invariants: odd dimensional Atiyah-Patodi-Singer problem. *Commun. Math. Phys.* 142 (1991), 139–168.

[KL] P. Kirk and M. Lesch, The η-invariant, Maslov index, and spectral flow for Dirac type operators on manifolds with boundary. *Preprint*, math.DG/0012123.

[MP] S. Minakshisundaram and A. Pleijel, Some properties of the eigenfunctions of the Laplace operator on Riemannian manifolds. *Canada J. Math.* 1 (1949), 242–256.

[MS] H. Mckean and I. M. Singer, Curvature and eigenvalues of the Laplacian. *J. Diff. Geom.* 1 (1967), 43–69.

[P] V. K. Patodi, Curvature and eigenforms of the Laplace operator. *J. Diff. Geom.* 5 (1971), 251–283.

[Yu] Y. Yu, *The Index Theorem and the Heat Equation Method.* Nankai Tracks in Mathematics Vol. 2. World Scientific, Singapore, 2001.

[Z] W. Zhang *Lectures on Chern-Weil Theory and Witten Deformations.* Nankai Tracks in Mathematics Vol. 4. World Scientific, Singapore, 2001.

Section 5. Topology

Mladen Bestvina: *The Topology of Out(F_n)* 373

Yu. V. Chekanov: *Invariants of Legendrian Knots* 385

M. Furuta: *Finite Dimensional Approximations in Geometry* 395

Emmanuel Giroux: *Géométrie de Contact: de la Dimension Trois vers les Dimensions Supérieures* 405

Lars Hesselholt: *Algebraic K-theory and Trace Invariants* 415

Eleny-Nicoleta Ionel: *Symplectic Sums and Gromov-Witten Invariants* 427

Peter Teichner: *Knots, von Neumann Signatures, and Grope Cobordism* 437

Ulrike Tillmann: *Strings and the Stable Cohomology of Mapping Class Groups* 447

Shicheng Wang: *Non-zero Degree Maps between 3-Manifolds* 457

ICM 2002 · Vol. II · 373–384

The Topology of $Out(F_n)$

Mladen Bestvina*

Abstract

We will survey the work on the topology of $Out(F_n)$ in the last 20 years or so. Much of the development is driven by the tantalizing analogy with mapping class groups. Unfortunately, $Out(F_n)$ is more complicated and less well-behaved.

Culler and Vogtmann constructed Outer Space X_n, the analog of Teichmüller space, a contractible complex on which $Out(F_n)$ acts with finite stabilizers. Paths in X_n can be generated using "foldings" of graphs, an operation introduced by Stallings to give alternative solutions for many algorithmic questions about free groups. The most conceptual proof of the contractibility of X_n involves folding.

There is a normal form of an automorphism, analogous to Thurston's normal form for surface homeomorphisms. This normal form, called a "(relative) train track map", consists of a cellular map on a graph and has good properties with respect to iteration. One may think of building an automorphism in stages, adding to the previous stages a building block that either grows exponentially or polynomially. A complicating feature is that these blocks are not "disjoint" as in Thurston's theory, but interact as upper stages can map over the lower stages.

Applications include the study of growth rates (a surprising feature of free group automorphisms is that the growth rate of f is generally different from the growth rate of f^{-1}), of the fixed subgroup of a given automorphism, and the proof of the Tits alternative for $Out(F_n)$. For the latter, in addition to train track methods, one needs to consider an appropriate version of "attracting laminations" to understand the dynamics of exponentially growing automorphisms and run the "ping-pong" argument. The Tits alternative is thus reduced to groups consisting of polynomially growing automorphisms, and this is handled by the analog of Kolchin's theorem (this is one instance where $Out(F_n)$ resembles $GL_n(\mathbb{Z})$ more than a mapping class group).

Morse theory has made its appearance in the subject in several guises. The original proof of the contractibility of X_n used a kind of "combinatorial" Morse function (adding contractible subcomplexes one at a time and studying the intersections). Hatcher-Vogtmann developed a "Cerf theory" for graphs. This is a parametrized version of Morse theory and it allows them to prove homological stability results. One can "bordify" Outer Space (by analogy with the Borel-Serre construction for arithmetic groups) to make the action

*Department of Mathematics, University of Utah, USA. E-mail: bestvina@math.utah.edu

of $Out(F_n)$ cocompact and then use Morse theory (with values in a certain ordered set) to study the connectivity at infinity of this new space. The result is that $Out(F_n)$ is a virtual duality group.

Culler-Morgan have compactified Outer Space, in analogy with Thurston's compactification of Teichmüller space. Ideal points are represented by actions of F_n on $\mathbb{R}$-trees. The work of Rips on group actions on $\mathbb{R}$-trees can be used to analyze individual points and the dynamics of the action of $Out(F_n)$ on the boundary. The topological dimension of the compactified Outer Space and of the boundary have been computed. The orbits in the boundary are not dense; however, there is a unique minimal closed invariant set. Automorphisms with irreducible powers act on compactified Outer Space with the standard North Pole – South Pole dynamics. By first finding fixed points in the boundary of Outer Space, one constructs a "hierarchical decomposition" of the underlying free group, analogous to the Thurston decomposition of a surface homeomorphism.

The geometry of Outer Space is not well understood. The most promising metric is not even symmetric, but this seems to be forced by the nature of $Out(F_n)$. Understanding the geometry would most likely allow one to prove rigidity results for $Out(F_n)$.

2000 Mathematics Subject Classification: 57M07, 20F65, 20E08.
Keywords and Phrases: Free group, Train tracks, Outer space.

1. Introduction

The aim of this note is to survey some of the topological methods developed in the last 20 years to study the group $Out(F_n)$ of outer automorphisms of a free group F_n of rank n. For an excellent and more detailed survey see also [69]. Stallings' paper [64] marks the turning point and for the earlier history of the subject the reader is referred to [55]. $Out(F_n)$ is defined as the quotient of the group $Aut(F_n)$ of all automorphisms of F_n by the subgroup of inner automorphisms. On one hand, abelianizing F_n produces an epimorphism $Out(F_n) \to Out(\mathbb{Z}^n) = GL_n(\mathbb{Z})$, and on the other hand $Out(F_n)$ contains as a subgroup the mapping class group of any compact surface with fundamental group F_n. A *leitmotiv* in the subject, promoted by Karen Vogtmann, is that $Out(F_n)$ satisfies a mix of properties, some inherited from mapping class groups, and others from arithmetic groups. The table below summarizes the parallels between topological objects associated with these groups.

Outer space is not a manifold and only a polyhedron, imposing a combinatorial character on $Out(F_n)$.

2. Stallings' Folds

A *graph* is a 1-dimensional cell complex. A map $f : G \to G'$ between graphs is *simplicial* if it maps vertices to vertices and open 1-cells homeomorphically to open

Mapping class groups	$Out(F_n)$	$GL_n(\mathbb{Z})$ **(arithmetic groups)**	**algebraic properties**
Teichmüller space	Culler-Vogtmann's Outer space	$GL_n(\mathbb{R})/O_n$ (symmetric spaces)	finiteness properties cohomological dimension
Thurston normal form	train track representative	Jordan normal form	growth rates fixed points (subgroups)
Harer's bordification	bordification of Outer space	Borel-Serre bordification	Bieri-Eckmann duality
measured laminations	$\mathbb{R}$-trees	flag manifold (Furstenberg boundary)	Kolchin theorem Tits alternative
Harvey's curve complex	?	Tits building	rigidity

1-cells. The simplicial map f is a *fold* if it is surjective and identifies two edges that share at least one vertex. A fold is a homotopy equivalence unless the two edges share both pairs of endpoints and in that case the induced homomorphism in π_1 corresponds to killing a basis element.

Theorem 1 (Stallings [63]). *A simplicial map $f : G \to G'$ between finite connected graphs can be factored as the composition*

$$G = G_0 \to G_1 \to \cdots \to G_k \to G'$$

where each $G_i \to G_{i+1}$ is a fold and $G_k \to G'$ is locally injective (an immersion). Moreover, such a factorization can be found by a (fast) algorithm.

In the absence of valence 1 vertices the last map $G_k \to G'$ can be thought of as the core of the covering space of G' corresponding to the image in π_1 of f. The following problems can be solved algorithmically using Theorem 1 (these were known earlier, but Theorem 1 provides a simple unified argument). Let F be a free group with a fixed finite basis.

- Find a basis of the subgroup H generated by a given finite collection $h_1, \cdots, h_k$ of elements of F.
- Given $w \in F$, decide if $w \in < h_1, \cdots, h_k >$.
- Given $w \in F$, decide if w is conjugate into $< h_1, \cdots, h_k >$.
- Given a homomorphism $\phi : F \to F'$ between two free groups of finite rank, decide if ϕ is injective, surjective.
- Given finitely generated $H < F$ decide if it has finite index.
- Given two f.g. subgroups $H_1, H_2 < F$ compute $H_1 \cap H_2$ and also the collection of subgroups $H_1 \cap H_2^g$ where $g \in F$. In particular, is H_1 malnormal?
- Represent a given automorphism of F as the composition of generators of $Aut(F)$ of the following form:
  Signed permutations: each a_i maps to a_i or to a_i^{-1}.
  Change of maximal tree: $a_1 \mapsto a_1$, $a_i \mapsto a_1^{\pm 1} a_i$ or $a_i \mapsto a_i a_1^{\pm 1}$ $(i > 1)$.
- Todd-Coxeter process [65].

3. Culler-Vogtmann's Outer space

Fix the wedge of n circles R_n and a natural identification $\pi_1(R_n) \cong F_n$ in which oriented edges correspond to the basis elements. Thus any $\phi \in Out(F_n)$ can be thought of as a homotopy equivalence $R_n \to R_n$. A *marked metric graph* is a pair (G, g) where

- G is a finite graph without vertices of valence 1 or 2.
- $g : R_n \to G$ is a homotopy equivalence (the *marking*).
- G is equipped with a path metric so that the sum of the lengths of all edges is 1.

Outer space X_n is the set of equivalence classes of marked metric graphs under the equivalence relation $(G, g) \sim (G', g')$ if there is an isometry $h : G \to G'$ such that gh and g' are homotopic [28].

If α is a loop in R_n we have the length function $l_\alpha : X_n \to \mathbb{R}$ where $l_\alpha(G, g)$ is the length of the immersed loop homotopic to $g(\alpha)$. The collection $\{l_\alpha\}$ as α ranges over all immersed loops in R_n defines an injection $X_n \to \mathbb{R}^\infty$ and the topology on X_n is defined so that this injection is an embedding. X_n naturally decomposes into open simplices obtained by varying edge-lengths on a fixed marked graph. The group $Out(F_n)$ acts on X_n on the right via

$$(G, g)\phi = (G, g\phi).$$

Theorem 2 (Culler-Vogtmann [28]). *X_n is contractible and the action of $Out(F_n)$ is properly discontinuous (with finite point stabilizers). X_n equivariantly deformation retracts to a $(2n-3)$-dimensional complex ($n > 1$).*

If (G, g) and (G', g') represent two points of X_n, there is a "difference of markings" map $h : G \to G'$ such that hg and g' are homotopic. Representing h as a composition of folds (appropriately interpreted) leads to a path in X_n from (G, g) to (G', g'). Arranging that these paths vary continuously with endpoints leads to a proof of contractibility of X_n [66],[60],[71].

Corollary 3. *The virtual cohomological dimension* $\mathrm{vcd}(Out(F_n)) = 2n-3$ *($n > 1$).*

Theorem 4 (Culler [26]). *Every finite subgroup of $Out(F_n)$ fixes a point of X_n.*

Outer space can be equivariantly compactified [27]. Points at infinity are represented by actions of F_n on $\mathbb{R}$-trees.

4. Train tracks

Any $\phi \in Out(F_n)$ can be represented as a cellular map $f : G \to G$ on a marked graph G. We say that ϕ is *reducible* if there is such a representative where

- G has no vertices of valence 1 or 2, and
- there is a proper f-invariant subgraph of G with at least one non-contractible component.

Otherwise, we say that ϕ is *irreducible.*

A cellular map $f : G \to G$ is a *train track map* if for every $k > 0$ the map $f^k : G \to G$ is locally injective on every open 1-cell. For example, homeomorphisms are train track maps and Culler's theorem guarantees that every $\phi \in Out(F_n)$ of finite order has a representative $f : G \to G$ which is a homeomorphism. More generally, we have

Theorem 5 (Bestvina-Handel [12]). *Every irreducible outer automorphism ϕ can be represented as a train track map $f : G \to G$.*

Any vertex $v \in G$ has a cone neighborhood, and the frontier points can be thought of as "germs of directions" at v. A train track map (or any cellular map that does not collapse edges) f induces the "derivative" map Df on these germs (on possibly different vertices). We declare two germs at the same vertex to be equivalent (and the corresponding "turn" *illegal*) if they get identified by some power of Df (and otherwise the turn is *legal*). An immersed loop in G is *legal* if every turn determined by entering and then exiting a vertex is legal. It follows that f sends legal loops to legal loops. This gives a method for computing the growth rate of ϕ, as follows. The *transition matrix* (a_{ij}) of f (or more generally of a cellular map $G \to G$ that is locally injective on edges) has a_{ij} equal to the number of times that the f-image of j^{th} edge crosses i^{th} edge. Applying the Perron-Frobenius theorem to the transition matrix, one can find a unique metric structure on G such that f expands lengths of edges (and also legal loops) by a factor $\lambda \geq 1$. For a conjugacy class γ in F_n the growth rate is defined as

$$GR(\phi, \gamma) = \limsup_{k \to \infty} \log(||\phi^k(\gamma)||)/k$$

where $||\gamma||$ is the word length of the cyclically reduced word representing γ. Growth rates can be computed using lengths of loops in G rather than in R_n.

Corollary 6. *If ϕ is irreducible as above, then either γ is a ϕ-periodic conjugacy class, or $GR(\phi, \gamma) = \log \lambda$. Moreover,* lim sup *can be replaced by* lim.

The proof of Theorem 5 uses a folding process that successively reduces the Perron-Frobenius number of the transition matrix until either a train track representative is found, or else a reduction of ϕ is discovered. This process is algorithmic (see [13],[21]).

Another application of train tracks is to fixed subgroups.

Theorem 7 (Bestvina-Handel [12]). *Let $\Phi : F_n \to F_n$ be an automorphism whose associated outer automorphism is irreducible. Then the fixed subgroup $Fix(\Phi)$ is trivial or cyclic. Without the irreducibility assumption, the rank of $Fix(\Phi)$ is at most n.*

It was known earlier by the work of Gersten [39] that $Fix(\Phi)$ has finite rank (for simpler proofs see [42],[25]). The last sentence in the above theorem was conjectured by Peter Scott. Subsequent work by Collins-Turner [24], Dicks-Ventura [31], Ventura [68], Martino-Ventura [58], imposed further restrictions on a subgroup of F_n that occurs as the fixed subgroup of an automorphism. To analyze reducible automorphisms, a more general version of a train track map is required.

Definition 8. *A cellular map $f : G \to G$ on a finite graph with no vertices of valence 1 that does not collapse any edges is a relative train track map if there is a filtration*

$$\emptyset = G_0 \subset \cdots \subset G_m = G$$

into f-invariant subgraphs with the following properties. Denote by H_r the closure of $G_r \setminus G_{r-1}$, and by M_r the part of the transition matrix corresponding to H_r. Then M_r is the zero matrix or an irreducible matrix. If M_r is irreducible and the Perron-Frobenius eigenvalue $\lambda_r > 1$ then:

- *the derivative Df maps the germs in H_r to germs in H_r,*
- *if α is a nontrivial path in G_{r-1} with endpoints in $G_{r-1} \cap H_r$ then $f(\alpha)$, after pulling tight, is also a nontrivial path with endpoints in $G_{r-1} \cap H_r$, and*
- *every legal path in H_r is mapped to a path that does not cross illegal turns in H_r.*

As an example, consider the automorphism $a \mapsto a, b \mapsto ab, c \mapsto caba^{-1}b^{-1}d$, $d \mapsto dbcd$ represented on the rose R_4. The strata are $\emptyset \subset G_1 = \{a\} \subset \{a, b\} \subset G$. H_1 and H_2 have $\lambda = 1$ while H_3 has $\lambda_3 > 1$. The following is an analog of Thurston's normal form for surface homeomorphisms.

Theorem 9. [12] *Every automorphism of F_n admits a relative train track representative.*

Consequently, automorphisms of F_n can be thought of as being built from building blocks (exponential and non-exponential kinds) but the later stages are allowed to map over the previous stages. This makes the study of automorphisms of F_n more difficult (and interesting) than the study of surface homeomorphisms. Other non-surface phenomena (present in linear groups) are:

- stacking up non-exponential strata produces (nonlinear) polynomial growth,
- the growth rate of an automorphism is generally different from the growth rate of its inverse.

5. Related spaces and structures

Unfortunately, relative train track representatives are far from unique. As a replacement, one looks for canonical objects associated to automorphisms that can be computed using relative train tracks. There are 3 kinds of such objects, all stemming from the surface theory: laminations, $\mathbb{R}$-trees, and hierarchical decompositions of F_n [59].

Laminations. Laminations were used in the proof of the Tits alternative for $Out(F_n)$. To each automorphism one associates finitely many attracting laminations. Each consists of a collection of "leaves", i.e. biinfinite paths in the graph G, or alternatively, of an F_n-orbit of pairs of distinct points in the Cantor set of ends of F_n. A leaf ℓ can be computed by iterating an edge in an exponentially growing stratum H_r. The other leaves are biinfinite paths whose finite subpaths appear as subpaths of ℓ. Some of the attracting laminations may be sublaminations of other

attracting laminations, and one focuses on the maximal (or *topmost*) laminations. It is possible to identify the basin of attraction for each such lamination. Let $\mathcal{H}$ be any subgroup of $Out(F_n)$. Some of the time it is possible to find to elements $f, g \in \mathcal{H}$ that attract each other's laminations and then the standard ping-pong argument shows that $< f, g > \cong F_2$. Otherwise, there is a finite set of attracting laminations permuted by $\mathcal{H}$, a finite index subgroup $\mathcal{H}_0 \subset \mathcal{H}$ that fixes each of these laminations and a homomorphism ("stretch factor") $\mathcal{H}_0 \to A$ to a finitely generated abelian group A whose kernel consists entirely of polynomially growing automorphisms. There is an analog of Kolchin's theorem that says that finitely generated groups of polynomially growing automorphisms can simultaneously be realized as relative train track maps on the same graph (the classical Kolchin theorem says that a group of unipotent matrices can be conjugated to be upper triangular, or equivalently that it fixes a point in the flag manifold). The main step in the proof of the analog of Kolchin's theorem is to find an appropriate fixed $\mathbb{R}$-tree in the boundary of Outer space. This leads to the Tits alternative for $Out(F_n)$:

Theorem 10 (Bestvina-Feighn-Handel [9],[10],[7]). *Any subgroup $\mathcal{H}$ of $Out(F_n)$ either contains F_2 or is virtually solvable.*

A companion theorem [8] (for a simpler proof see [1]) is that solvable subgroups of $Out(F_n)$ are virtually abelian.

$\mathbb{R}$-trees. Points in the compactified Outer space are represented as F_n-actions on $\mathbb{R}$-trees. It is then not surprising that the Rips machine [5], which is used to understand individual actions, provides a new tool to be deployed to study $Out(F_n)$. Gaboriau, Levitt, and Lustig [37] and Sela [59] find another proof of Theorem 7. Gaboriau and Levitt compute the topological dimension of the boundary of Outer Space [36]. Levitt and Lustig show [51] that automorphisms with irreducible powers have the standard north-south dynamics on the compactified Outer space. Guirardel [43] shows that the action of $Out(F_n)$ on the boundary does not have dense orbits; however, there is a unique minimal closed invariant set. For other applications of $\mathbb{R}$-trees in geometric group theory, the reader is referred to the survey [2].

Cerf theory. An advantage of $Aut(F_n)$ over $Out(F_n)$ is that there is a natural inclusion $Aut(F_n) \to Aut(F_{n+1})$. One can define *Auter Space* AX_n similarly to Outer space, except that all graphs are equipped with a base vertex, which is allowed to have valence 2. The degree of the base vertex v is $2n - \text{valence}(v)$. Denote by D_n^k the subcomplex of AX_n consisting of graphs of degree $\leq k$. Hatcher-Vogtmann [47] develop a version of Cerf theory and show that D_n^k is $(k-1)$-connected. Since the quotient $D_n^k/Aut(F_n)$ stabilizes when n is large, one sees that (rational) homology $H_i(Aut(F_n))$ also stabilizes when n is large ($n \geq 3i/2$). Hatcher-Vogtmann show that the same is true for integral homology and in the range $n \geq 2i+3$. They also make explicit computations in low dimensions [49] and all stable rational homology groups H_i vanish for $i \leq 7$.

Bordification. The action of $Out(F_n)$ on Outer space X_n is not cocompact. By analogy with Borel-Serre bordification of symmetric spaces [14] and Harer's bordification of Teichmüller space [44], Bestvina and Feighn [6] bordify X_n, i.e. equivariantly add ideal points so that the action on the new space BX_n is cocompact. This is done by separately compactifying every simplex with missing faces in X_n

and then gluing these together. To see the idea, consider the case of the theta-graph in rank 2. Varying metrics yields a 2-simplex σ without the vertices. As a sequence of metrics approaches a missing vertex, the lengths of two edges converge to 0. Restricting a metric to these two edges and normalizing so that the total length is 1 gives a point in $[0,1]$ (the length of one of the edges), and a way to compactify σ by adding an interval for each missing vertex. The compactified σ is a hexagon. This procedure equips the limiting theta graph with a metric that may vanish on two edges, in which case a "secondary metric" is defined on their union. In general, a graph representing a point in the bordification is equipped with a sequence of metrics, each defined on the core of the subgraph where the previous metric vanishes.

Lengths of curves (at various scales) provide a "Morse function" on BX_n with values in a product of $[0,\infty)$'s with the target lexicographically ordered. The sublevel and superlevel sets intersect each cell in a semi-algebraic set and it is possible to study how the homotopy types change as the level changes. A distinct advantage of BX_n over the spine of X_n (an equivariant deformation retract) is that the change in homotopy type of superlevel sets as the level decreases is very simple – via attaching of cells of a fixed dimension.

Theorem 11 (Bestvina-Feighn [6]). *BX_n and $Out(F_n)$ are $(2n-5)$-connected at infinity, and $Out(F_n)$ is a virtual duality group of dimension $2n-3$.*

Mapping tori. If $\phi : F_n \to F_n$ is an automorphism, form the mapping torus $M(\phi)$. This is the fundamental group of the mapping torus $G \times [0,1]/(x,1) \sim (f(x),0)$ of any representative $f : G \to G$, and it plays the role analogous to 3-manifolds that fiber over the circle. Such a group is always coherent [33]. A quasi-isometry classification of these groups seems out of reach, but the following is known. When ϕ has no periodic conjugacy classes, $M(\phi)$ is a hyperbolic group [20]. When ϕ has polynomial growth, $M(\phi)$ satisfies quadratic isoperimetric inequality [57] and moreover, $M(\phi)$ quasi-isometric to $M(\psi)$ for ψ growing polynomially forces ψ to grow as a polynomial of the same degree [56]. Bridson and Groves announced [16] that $M(\phi)$ satisfies quadratic isoperimetric inequality for all ϕ.

Geometry. Perhaps the biggest challenge in the field is to find a good geometry that goes with $Out(F_n)$. The payoff would most likely include rigidity theorems for $Out(F_n)$. Both mapping class groups and arithmetic groups act isometrically on spaces of nonpositive curvature. Unfortunately, the results to date for $Out(F_n)$ are negative. Bridson [15] showed that Outer space does not admit an equivariant piecewise Euclidean $CAT(0)$ metric. $Out(F_n)$ $(n > 2)$ is far from being $CAT(0)$ [17],[40].

An example of a likely rigidity theorem is that higher rank lattices in simple Lie groups do not embed into $Out(F_n)$. A possible strategy is to follow the proof in [11] of the analogous fact for mapping class groups. The major missing piece of the puzzle is the replacement for Harvey's curve complex; a possible candidate is described in [48].

References

[1] Emina Alibegović, *Translation lengths in $Out(F_n)$*, 2000, to appear.

[2] Mladen Bestvina, **R**-*trees in topology, geometry, and group theory*, Handbook of geometric topology (R.J. Daverman and R.B. Sher, eds.), Elsevier Science B.V., 2002.

[3] Mladen Bestvina and Noel Brady, *Morse theory and finiteness properties of groups*, Invent. Math. **129** (1997), no. 3, 445–470.

[4] Mladen Bestvina and Mark Feighn, *A combination theorem for negatively curved groups*, J. Differential Geom. **35** (1992), no. 1, 85–101, Addendum and correction: **43** (1996), no. 4, 783–788.

[5] ———, *Stable actions of groups on real trees*, Invent. Math. **121** (1995), no. 2, 287–321.

[6] ———, *The topology at infinity of $Out(F_n)$*, Invent. Math. **140** (2000), no. 3, 651–692.

[7] Mladen Bestvina, Mark Feighn, and Michael Handel, *The Tits Alternative for $Out(F_n)$ II: a Kolchin Theorem*, to appear, 1996.

[8] ———, *The Tits Alternative for $Out(F_n)$ III: Solvable Subgroups*, preprint, 1996.

[9] ———, *Laminations, trees, and irreducible automorphisms of free groups*, Geom. Funct. Anal. **7** (1997), no. 2, 215–244, Erratum: **7** (1997), no. 6, 1143.

[10] ———, *The Tits alternative for* $\mathrm{Out}(F_n)$. *I. Dynamics of exponentially-growing automorphisms*, Ann. of Math. (2) **151** (2000), no. 2, 517–623. MR 2002a:20034

[11] Mladen Bestvina and Koji Fujiwara, *Bounded cohomology of subgroups of mapping class groups*, Geom. Top. **6** (2002), 69–89.

[12] Mladen Bestvina and Michael Handel, *Train tracks and automorphisms of free groups*, Ann. of Math. (2) **135** (1992), no. 1, 1–51.

[13] ———, *Train-tracks for surface homeomorphisms*, Topology **34** (1995), no. 1, 109–140.

[14] A. Borel and J.-P. Serre, *Corners and arithmetic groups*, Comment. Math. Helv. **48** (1973), 436–491, Avec un appendice: Arrondissement des variétés à coins, par A. Douady et L. Hérault.

[15] Martin R. Bridson, *Geodesics and curvature in metric simplicial complexes*, Group theory from a geometrical viewpoint (Trieste, 1990), World Sci. Publishing, River Edge, NJ, 1991, 373–463.

[16] Martin R. Bridson and Daniel Groves, *The quadratic isoperimetric inequality for mapping tori of free group automorphisms I: Positive automorphisms*, preprint.

[17] Martin R. Bridson and Karen Vogtmann, *On the geometry of the automorphism group of a free group*, Bull. London Math. Soc. **27** (1995), no. 6, 544–552.

[18] ———, *Automorphisms of automorphism groups of free groups*, J. Algebra **229** (2000), no. 2, 785–792.

[19] ———, *The symmetries of Outer space*, Duke Math. J. **106** (2001), no. 2, 391–409.

[20] Peter Brinkmann, *Hyperbolic automorphisms of free groups*, Geom. Funct.

Anal. **10** (2000), no. 5, 1071–1089.
[21] ———, *An implementation of the Bestvina-Handel algorithm for surface homeomorphisms*, Experiment. Math. **9** (2000), no. 2, 235–240.
[22] Marshall M. Cohen and Martin Lustig, *On the dynamics and the fixed subgroup of a free group automorphism*, Invent. Math. **96** (1989), no. 3, 613–638.
[23] ———, *Very small group actions on* **R***-trees and Dehn twist automorphisms*, Topology **34** (1995), no. 3, 575–617.
[24] D. J. Collins and E. C. Turner, *All automorphisms of free groups with maximal rank fixed subgroups*, Math. Proc. Cambridge Philos. Soc. **119** (1996), no. 4, 615–630.
[25] Daryl Cooper, *Automorphisms of free groups have finitely generated fixed point sets*, J. Algebra **111** (1987), no. 2, 453–456.
[26] Marc Culler, *Finite groups of outer automorphisms of a free group*, Contributions to group theory, Amer. Math. Soc., Providence, RI, 1984, 197–207.
[27] Marc Culler and John W. Morgan, *Group actions on* **R***-trees*, Proc. London Math. Soc. (3) **55** (1987), no. 3, 571–604.
[28] Marc Culler and Karen Vogtmann, *Moduli of graphs and automorphisms of free groups*, Invent. Math. **84** (1986), no. 1, 91–119.
[29] ———, *The boundary of Outer space in rank two*, Arboreal group theory (Berkeley, CA, 1988), Springer, New York, 1991, 189–230.
[30] ———, *A group-theoretic criterion for property* FA, Proc. Amer. Math. Soc. **124** (1996), no. 3, 677–683.
[31] Warren Dicks and Enric Ventura, *Irreducible automorphisms of growth rate one*, J. Pure Appl. Algebra **88** (1993), no. 1-3, 51–62.
[32] ———, *The group fixed by a family of injective endomorphisms of a free group*, American Mathematical Society, Providence, RI, 1996.
[33] Mark Feighn and Michael Handel, *Mapping tori of free group automorphisms are coherent*, Ann. of Math. (2) **149** (1999), no. 3, 1061–1077.
[34] Edward Formanek and Claudio Procesi, *The automorphism group of a free group is not linear*, J. Algebra **149** (1992), no. 2, 494–499.
[35] Damien Gaboriau, Andre Jaeger, Gilbert Levitt, and Martin Lustig, *An index for counting fixed points of automorphisms of free groups*, Duke Math. J. **93** (1998), no. 3, 425–452.
[36] Damien Gaboriau and Gilbert Levitt, *The rank of actions on* **r***-trees*, Ann. Sci. École Norm. Sup. (4) **28** (1995), no. 5, 549–570.
[37] Damien Gaboriau, Gilbert Levitt, and Martin Lustig, *A dendrological proof of the Scott conjecture for automorphisms of free groups*, Proc. Edinburgh Math. Soc. (2) **41** (1998), no. 2, 325–332.
[38] S. M. Gersten, *A presentation for the special automorphism group of a free group*, J. Pure Appl. Algebra **33** (1984), no. 3, 269–279.
[39] ———, *Fixed points of automorphisms of free groups*, Adv. in Math. **64** (1987), no. 1, 51–85.
[40] ———, *The automorphism group of a free group is not a* CAT(0) *group*, Proc. Amer. Math. Soc. **121** (1994), no. 4, 999–1002.
[41] S. M. Gersten and J. R. Stallings, *Irreducible outer automorphisms of a free*

group, Proc. Amer. Math. Soc. **111** (1991), no. 2, 309–314.
[42] Richard Z. Goldstein and Edward C. Turner, *Fixed subgroups of homomorphisms of free groups*, Bull. London Math. Soc. **18** (1986), no. 5, 468–470.
[43] Vincent Guirardel, *Dynamics of* $\mathrm{out}(F_n)$ *on the boundary of Outer space*, Ann. Sci. École Norm. Sup. (4) **33** (2000), no. 4, 433–465.
[44] John L. Harer, *The virtual cohomological dimension of the mapping class group of an orientable surface*, Invent. Math. **84** (1986), no. 1, 157–176.
[45] W. J. Harvey, *Geometric structure of surface mapping class groups*, Homological group theory (Proc. Sympos., Durham, 1977), Cambridge Univ. Press, Cambridge, 1979, 255–269.
[46] Allen Hatcher and Karen Vogtmann, *Isoperimetric inequalities for automorphism groups of free groups*, Pacific J. Math. **173** (1996), no. 2, 425–441.
[47] ———, *Cerf theory for graphs*, J. London Math. Soc. (2) **58** (1998), no. 3, 633–655.
[48] ———, *The complex of free factors of a free group*, Quart. J. Math. Oxford Ser. (2) **49** (1998), no. 196, 459–468.
[49] ———, *Rational homology of* $\mathrm{Aut}(F_n)$, Math. Res. Lett. **5** (1998), no. 6, 759–780.
[50] Gilbert Levitt, *Non-nesting actions on real trees*, Bull. London Math. Soc. **30** (1998), no. 1, 46–54.
[51] Gilbert Levitt and Martin Lustig, *Irreducible automorphisms of* F_n *have north-south dynamics on compactified Outer space*, preprint 2002.
[52] ———, *Periodic ends, growth rates, Hölder dynamics for automorphisms of free groups*, Comment. Math. Helv. **75** (2000), no. 3, 415–429.
[53] Martin Lustig, *Structure and conjugacy for automorphisms of free groups I*, Max Planck Institute preprint MPI 2000 - 130.
[54] ———, *Structure and conjugacy for automorphisms of free groups II*, Max Planck Institute preprint MPI 2001 - 4.
[55] Roger C. Lyndon and Paul E. Schupp, *Combinatorial group theory*, Springer-Verlag, Berlin, 1977, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 89.
[56] Nataša Macura, *Detour functions and quasi-isometries*, to appear.
[57] ———, *Quadratic isoperimetric inequality for mapping tori of polynomially growing automorphisms of free groups*, Geom. Funct. Anal. **10** (2000), no. 4, 874–901.
[58] Armando Martino and Enric Ventura, *On automorphism-fixed subgroups of a free group*, J. Algebra **230** (2000), no. 2, 596–607.
[59] Zlil Sela, *The Nielsen-Thurston classification and automorphisms of a free group. I*, Duke Math. J. **84** (1996), no. 2, 379–397.
[60] Richard Skora, *Deformations of length functions in groups*, preprint 1989.
[61] John Smillie and Karen Vogtmann, *A generating function for the Euler characteristic of* $\mathrm{Out}(F_n)$, Proceedings of the Northwestern conference on cohomology of groups (Evanston, Ill., 1985), vol. 44, 1987, 329–348.
[62] ———, *Length functions and Outer space*, Michigan Math. J. **39** (1992), no. 3, 485–493.

[63] John R. Stallings, *Topology of finite graphs*, Invent. Math. **71** (1983), no. 3, 551–565.

[64] ———, *Finite graphs and free groups*, Combinatorial methods in topology and algebraic geometry (Rochester, N.Y., 1982), Amer. Math. Soc., Providence, RI, 1985, 79–84.

[65] John R. Stallings and A. Royce Wolf, *The Todd-Coxeter process, using graphs*, Combinatorial group theory and topology (Alta, Utah, 1984), Princeton Univ. Press, Princeton, NJ, 1987, 157–161.

[66] Michael Steiner, *Gluing data and group actions on* **R***-trees*, Thesis, Columbia University, 1988.

[67] Edward C. Turner, *Finding indivisible Nielsen paths for a train track map*, Combinatorial and geometric group theory (Edinburgh, 1993), Cambridge Univ. Press, Cambridge, 1995, 300–313.

[68] Enric Ventura, *On fixed subgroups of maximal rank*, Comm. Algebra **25** (1997), no. 10, 3361–3375.

[69] Karen Vogtmann, *Automorphisms of free groups and Outer space*, to appear.

[70] ———, *Local structure of some* $\mathrm{Out}(F_n)$*-complexes*, Proc. Edinburgh Math. Soc. (2) **33** (1990), no. 3, 367–379.

[71] Tad White, *Fixed points of finite groups of free group automorphisms*, Proc. Amer. Math. Soc. **118** (1993), no. 3, 681–688.

ICM 2002 · Vol. II · 385–394

Invariants of Legendrian Knots

Yu. V. Chekanov*

Abstract

We present two different constructions of invariants for Legendrian knots in the standard contact space $\mathbb{R}^3$. These invariants are defined combinatorially, in terms of certain planar projections, and are useful in distinguishing Legendrian knots that have the same classical invariants but are not Legendrian isotopic.

2000 Mathematics Subject Classification: 57R17.
Keywords and Phrases: Legendrian submanifold, Legendrian knot.

1. Introduction

1.1. Legendrian knots

A smooth knot L in the standard contact space $(\mathbb{R}^3, \alpha) = (\{(q,p,u)\}, du - pdq)$ is called Legendrian if it is everywhere tangent to the 2-plane distribution $\ker(\alpha)$ (or, in other words, if the restriction of α to L vanishes). Two Legendrian knots are Legendrian isotopic if the they can be connected by a smooth path in the space of Legendrian knots (or, equivalently, if one can be sent to another by a diffeomorphism g of $\mathbb{R}^3$ such that $g^*\alpha = \varphi\alpha$, where $\varphi > 0$). In order to visualize a knot in $\mathbb{R}^3$, it is convenient to project it to a plane. In the Legendrian case, the character of the resulting picture will depend on the choice of the projection. The useful two are: the Lagrangian projection $\pi\colon \mathbb{R}^3 \to \mathbb{R}^2$, $(q,p,u) \mapsto (q,p)$, and the front projection $\sigma\colon \mathbb{R}^3 \to \mathbb{R}^2$, $(p,q,u) \mapsto (q,u)$. In Figure 1, two projections of the simplest Legendrian knot (unknot) are shown.

We say that a Legendrian knot $L \subset \mathbb{R}^3$ is π-generic if all self-intersections of the immersed curve $\pi(L)$ are transverse double points. We can represent a π-generic Legendrian knot L by its (Lagrangian) diagram: the curve $\pi(L) \subset \mathbb{R}^2$, at every crossing of which the overpassing branch (the one with the greater value of u) is marked. Of course, not every abstract knot diagram in $\mathbb{R}^2$ is a diagram of

*Moscow Center for Continuous Mathematical Education, B. Vlasievsky per. 11, Moscow 119002, Russia. E-mail: chekanov@mccme.ru

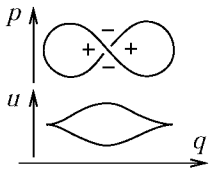


Figure 1: Lagrangian projection and front projection

a Legendrian knot, or is oriented diffeomorphic to such (it requires a bit of extra work to check whether a given diagram corresponds to a Legendrian knot, cf. [1]).

Given a Legendrian knot $L \subset \mathbb{R}^3$, its σ-projection, or front, $\sigma(L) \subset \mathbb{R}^2$ is a singular curve with nowhere vertical tangent vectors. Its singularities, generically, are semi-cubic cusps and transverse double points. We say that L is σ-generic if, moreover, all self-intersections of $\sigma(L)$ have different q-coordinates. Every closed planar curve with these types of singularities and nowhere vertical tangent vectors is a front of a Legendrian knot. Note that there is no need to explicitly indicate the type of a crossing of a front: the overpassing branch (the one with the greater value of p) is always the one with the greater slope.

1.2. Classical invariants

The so-called classical invariants of an oriented Legendrian knot L are defined as follows. The first of them is, formally, just the smooth isotopy type of L. The Thurston–Bennequin number $\beta(L)$ of L is the linking number (with respect to the orientation defined by $\alpha \wedge d\alpha$) between L and $s(L)$, where s is a small shift along the u direction. The Maslov number $m(L)$ (which actually is an invariant of Legendrian immersion) is twice the rotation number of the projection of L to the (q,p) plane (or, equivalently, the value of the Maslov 1-cohomology class on the fundamental class of L). The change of orientation on L changes the sign of $m(L)$ and preserves $\beta(L)$. The Thurston-Bennequin number of a π-generic Legendrian knot L can be computed by counting the crossings of its Lagrangian diagram $\pi(L)$ with signs:

$$\beta(L) = \#(\times) - \#(\times)$$

(where the q axis is horizontal and the p axis is vertical). In terms of the front projection, the classical invariants can be computed as follows. The Maslov number of a σ-generic oriented Legendrian knot L is the number of the right cusps of the front $\sigma(L)$, counted with signs depending on the orientations:

$$m(L) = \#(\rightarrow) - \#(\rightarrow).$$

The Thurston-Bennequin number of L is the number of crossings of $\sigma(L)$ counted with signs minus half the total number of cusps (= the number of right cusps):

$$\beta(L) = \#(\times) + \#(\times) - \#(\times) - \#(\times) - \#(\rightarrow).$$

For the Legendrian knot shown in Figure 1, we have $m = 0$, $\beta = -1$.

1.3. New invariants and classification results

It is easy to show that every smooth knot admits a Legendrian realization. A natural question to ask is whether there exists a pair of Legendrian knots which have the same classical invariants but are not Legendrian isotopic. The answer is positive:

Theorem 1.1. [1, 2] *There exist Legendrian knots L, L' (see Figure 2 on p. 390, Figure 6 on p. 393) that have the same classical invariants (smooth knot type 5_2, $m = 0$, $\beta = 1$) but are not Legendrian isotopic.*

In the next two sections, we present two combinatorial constructions of Legendrian knot invariants. The first one associates to the Lagrangian projection of a Legendrian knot a differential graded algebra (DGA). The second construction deals with decompositions of the front projection into closed curves. Each of the two provides a proof of Theorem 1.1. The invariants do not change when the orientation of the knot reverses, so essentially they are invariants of non-oriented Legendrian knots. It should be mentioned that these constructions also produce, with minor modifications, invariants of Legendrian links.

The number of Legendrian knots with given classical invariants is known to be finite [3]. Eliashberg and Fraser gave a classification of Legendrian realization for smooth unknots [5, 6]. It turned out that smooth unknots are Legendrian simple in the sense that the Legendrian isotopy types of their Legendrian realizations are determined by the classical invariants. Etnyre and Honda obtained a classification of Legendrian realization for torus knots and the figure eight knot [8]. Again, these smooth knot types proved to be Legendrian simple. The 5_2 is the simplest knot type for which the classification is not known. By Theorem 1.1, the type 5_2 is not Legendrian simple. Conjecturally, two Legendrian knots of smooth type 5_2 with the same classical invariants are Legendrian isotopic unless they form the pair L, L' from Theorem 1.1. Several interesting examples of knots with coinciding classical invariants but not Legendrian isotopic were constructed by Ng [14, 15].

2. DGA of a Legendrian knot

2.1. Definitions

In this section, we associate with every π-generic Legendrian knot L a DGA (A, ∂) over $\mathbb{Z}/2\mathbb{Z}$ ([1]; a similar construction was also given by Eliashberg). This DGA is related to the symplectic field theory introduced by Eliashberg, Givental, and Hofer in [7] (see [10]).

Let $\{a_1, \dots, a_n\}$ be the set of crossings of $Y = \pi(L)$. Define A to be the tensor algebra (free associative unital algebra) $T(a_1, \dots, a_n)$ with generators $a_1, \dots, a_n$. The grading on A takes values in the group $\mathbb{Z}/m(L)\mathbb{Z}$ and is defined as follows. Given a crossing a_j, consider the points $z_+, z_- \in L$ such that $\pi(z_+) = \pi(z_-) = a_j$ and the u-coordinate of z_+ is greater than the u-coordinate of z_-. These points divide L into two pieces, γ_1 and γ_2, which we orient from z_+ to z_-. We can assume,

without loss of generality, that the intersecting branches are orthogonal at a. Then, for $\varepsilon \in \{1,2\}$, the rotation number of the curve $\pi(\gamma_\varepsilon)$ has the form $N_\varepsilon/2 + 1/4$, where $N_\varepsilon \in \mathbb{Z}$. Clearly, $N_1 - N_2$ is equal to $\pm m(L)$. Hence N_1 and N_2 represent the same element of the group $\Gamma = \mathbb{Z}/m(L)\mathbb{Z}$, which we define to be the degree of a_j.

We are going to define the differential ∂. For every natural k, fix a (curved) convex k-gon $\Pi_k \subset \mathbb{R}^2$ whose vertices $x_0^k, \dots, x_{k-1}^k$ are numbered counter-clockwise. The form $dq \wedge dp$ defines an orientation on $\mathbb{R}^2$. Denote by $W_k(Y)$ the collection of smooth orientation-preserving immersions $f\colon \Pi_k \to \mathbb{R}^2$ such that $f(\partial\Pi_k) \subset Y$. Note that $f \in W_k(Y)$ implies $f(x_i^k) \in \{a_1, \dots, a_n\}$. Consider the set of nonparametrized immersions $\widetilde{W}_k(Y)$, which is the quotient of $W_k(Y)$ by the action of the group $\{g \in \mathrm{Diff}_+(\Pi_k) | g(x_i^k) = x_i^k\}$. The diagram Y divides a neighbourhood of each of its crossings into four sectors. We call positive two of them which are swept out by the underpassing curve rotating counter-clockwise, and negative the other two (the sectors are marked in Figure 1). For each vertex x_i^k of the polygon Π_k, a smooth immersion $f \in \widetilde{W}_k(Y)$ maps its neighbourhood in Π_k to either a positive or a negative sector; we shall say that x_i^k is, respectively, a positive or a negative vertex for f. Define the set $W_k^+(Y)$ to consist of immersions $f \in \widetilde{W}_k(Y)$ such that the vertex x_0^k is positive for f, and all other vertices are negative. Let $W_k^+(Y, a_j) = \{f \in W_k^+(Y) \mid f(x_0^k) = a_j\}$. Denote $A_1 = \{a_1, \dots, a_n\} \otimes \mathbb{Z}/2\mathbb{Z} \subset A$, $A_k = (A_1)^{\otimes k}$. Then $A = \oplus_{l=0}^{\infty} A_l$. Let $\partial = \sum_{k \geq 0} \partial_k$, where $\partial_k(A_i) \in A_{i+k-1}$. Define

$$\partial_k(a_j) = \sum_{f \in W_{k+1}^+(Y, a_j)} f(x_1) \cdots f(x_k)$$

(for $k = 0$, we have $\partial_0(a_j) = \#(W_1^+(Y, a_j))$, and extend ∂ to A by linearity and the Leibniz rule. The following theorem says that (A, ∂) is indeed a DGA:

Theorem 2.1. *The differential ∂ is well defined. We have* $\deg(\partial) = -1$ *and* $\partial^2 = 0$.

Define the (l-th, where $l \in \Gamma$) stabilization of a DGA $(T(a_1, \dots, a_n), \partial)$ to be the DGA $(T(a_1, \dots, a_n, a_{n+1}, a_{n+2}), \partial)$, where $\deg(a_{n+1}) = l$, $\deg(a_{n+2}) = l - 1$, $\partial(a_{n+1}) = a_{n+2}$, and ∂ acts on $a_1, \dots, a_n$ as before. An automorphism of $T(a_1, \dots, a_n)$ is called elementary if it sends a_i to $a_i + v$, where v does not involve a_i, and fixes a_j for $j \neq i$. Two DGAs $(T(a_1, \dots, a_n), \partial)$, $(T(a_1, \dots, a_n), \partial')$ are called tame isomorphic if one can be obtained from another by a composition of elementary automorphisms; they are called stable tame isomorphic if they become tame isomorphic after (iterated) stabilizations.

Theorem 2.2. *Let (A, ∂), (A', ∂') be the DGAs of (π-generic) Legendrian knots L, L'. If L and L' are Legendrian isotopic then (A, ∂) and (A', ∂') are stable tame isomorphic. In particular, the homology rings $H(A, \partial) = \ker(\partial)/\mathrm{im}(\partial)$ and $H(A', \partial') = \ker(\partial)/\mathrm{im}(\partial)$ are isomorphic as graded rings.*

The hard part in the proof of Theorem 2.1 is to show that $\partial^2 = 0$. The proof of this fact mimics, in a combinatorial way, the classical gluing–compactness argument of the Floer theory (cf. [11]). The proof of Theorem 2.2 involves a careful study of

the behaviour of the DGA associated with a Legendrian knot when its Lagrangian diagram goes through elementary bifurcations (Legendrian Reidemeister moves).

It turns out that one cannot replace the coefficient ring $\mathbb{Z}/2\mathbb{Z}$ by $\mathbb{Z}$: in some sense, our homology theory is not oriented. However, the construction described above can be modified to associate with a Legendrian knot L a DGA graded by $\mathbb{Z}$ and having $\mathbb{Z}[s,s^{-1}]$ (where $\deg(s) = m(L)$) as a coefficient ring [10]. After reducing the grading to $\mathbb{Z}/m(L)\mathbb{Z}$, and applying the homomorphism $\mathbb{Z}[s,s^{-1}] \to \mathbb{Z}/2\mathbb{Z}$ sending both s and $1 \in \mathbb{Z}$ to $1 \in \mathbb{Z}/2\mathbb{Z}$, this $\mathbb{Z}[s,s^{-1}]$-DGA becomes the $\mathbb{Z}/2\mathbb{Z}$-DGA of the knot L.

2.2. Poincaré polynomials

Homology rings of DGAs can be hard to work with. We are going to define an easily computable invariant I, which is a finite subset of the group monoid $\mathbb{N}_0[\Gamma]$, where $\mathbb{N}_0 = \{0,1,\ldots\}$, $\Gamma = \mathbb{Z}/m(L)\mathbb{Z}$. Assume that $\partial_0 = 0$. Then $\partial_1^2 = 0$. Since $\partial(A_1) \subset A_1$, we can consider the homology $H(A_1,\partial_1) = \ker(\partial_1|_{A_1})/\mathrm{im}\,(\partial_1|_{A_1})$, which is a vector space graded by the cyclic group Γ. Define the Poincaré polynomial $P_{(A,\partial)} \in \mathbb{N}_0[\Gamma]$ by

$$P_{(A,\partial)}(t) = \sum_{\lambda\in\Gamma} \dim\big(H_\lambda(A_1,\partial_1)\big)\, t^\lambda,$$

where $H_\lambda(A_1,\partial_1)$ is the degree λ homogeneous component of $H(A_1,\partial_1)$. Define the group $\mathrm{Aut}_0(A)$ to consist of graded automorphisms of A such that for each $i \in \{1,\ldots,n\}$ we have $g(a_i) = a_i + c_i$, where $c_i \in A_0 = \mathbb{Z}/2\mathbb{Z}$. (of course, $c_i = 0$ when $\deg(a_i) \neq 0$). Consider the set $U_0(A,\partial)$ consisting of automorphisms $g \in \mathrm{Aut}_0(A)$ such that $(\partial^g)_0 = 0$ (where $\partial^g = g^{-1} \circ \partial \circ g$). Define

$$I(A,\partial) = \{\, P_{(A,\partial^g)} \mid g \in U_0(A,\partial) \,\}.$$

Since $\mathrm{Aut}_0(A)$ has at most 2^n elements, this invariant is not hard to compute. We can associate with every (π-generic) Legendrian knot L the set $I(L) = I(A_L,\partial_L)$. Note that $P(-1) = \beta(L)$ for $P \in I(L)$. One can show that I is an invariant of stable tame DGA isomorphism. Hence Theorem 2.2 implies the following

Corollary 2.3. *If L is Legendrian isotopic to L' then $I(L) = I(L')$.*

The set $I(L)$ can be empty (cf. Section 4) but no examples are known where $I(L)$ contains more than one element. Also, for all known examples of pairs L, L' of Legendrian knots with coinciding classical invariants we have $P(1) = P'(1)$, where $P \in I(L)$, $P' \in I(L')$. Other, more complicated invariants of stable tame isomorfism were developed and applied to distinguishing Legendrian knots in [15].

2.3. Examples

1a. Let $(A,\partial) = (T(a_1,\ldots,a_9),\partial)$ be the DGA of the Legendrian knot L given in Figure 2. We have $m(L) = 0$, $\beta(L) = 1$, $\deg(a_i) = 1$ for $i \leq 4$, $\deg(a_5) = 2$, $\deg(a_6) = -2$, $\deg(a_i) = 0$ for $i \geq 7$, $\partial(a_1) = 1 + a_7 + a_7a_6a_5$, $\partial(a_2) = 1 + a_9 + a_5a_6a_9$, $\partial(a_3) = 1 + a_8a_7$, $\partial(a_4) = 1 + a_8a_9$, $\partial(a_i) = 0$ for $i \geq 5$.

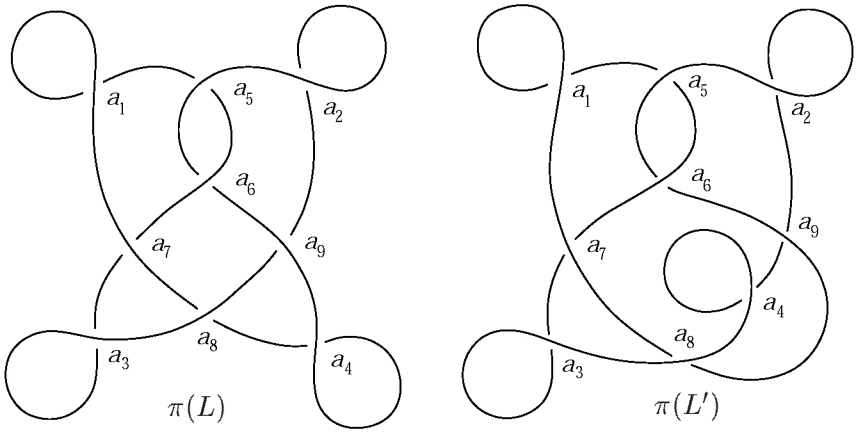


Figure 2: Lagrangian projections of two Legendrian 5_2 knots

1b. Let $(A',\partial) = (T(a_1,\dots,a_9),\partial)$ be the DGA of the Legendrian knot L' given in Figure 2. We have $m(L') = 0$, $\beta(L') = 1$, $\deg(a_i) = 1$ for $i \le 4$, $\deg(a_i) = 0$ for $i \ge 5$, $\partial(a_1) = 1 + a_7 + a_5 + a_7a_6a_5 + a_9a_8a_5$, $\partial(a_2) = 1 + a_9 + a_5a_6a_9$, $\partial(a_3) = 1 + a_8a_7$, $\partial(a_4) = 1 + a_8a_9$, $\partial(a_i) = 0$ for $i \ge 5$.

An explicit computation shows that $I(L) = \{t^{-2} + t^1 + t^2\}$, $I(L') = \{2t^0 + t^1\}$ and hence Theorem 1.1 follows from Corollary 2.3.

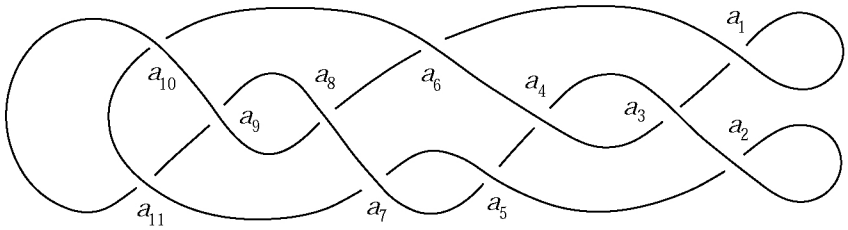


Figure 3: Lagrangian projection of a Legendrian 6_2 knot

2. [14] Let $(A,\partial) = (T(a_1,\dots,a_{11}),\partial)$ be the DGA of the Legendrian knot K given in Figure 3. We have $m(K) = 0$, $\beta(K) = -7$, $\deg(a_i) = 1$ for $i \in \{1,2,7,9,10\}$, $\deg(a_i) = 0$ for $i \in \{3,4\}$, $\deg(a_i) = -1$ for $i \in \{5,6,8,11\}$; $\partial(a_1) = 1 + a_{10}a_5a_3$, $\partial(a_2) = 1 + a_3 + a_3a_6a_{10} + a_3a_{11}a_7$, $\partial(a_4) = a_5 + a_{11} + a_{11}a_7a_5$, $\partial(a_6) = a_{11}a_8$, $\partial(a_7) = a_8a_{10}$, $\partial(a_9) = 1 + a_{10}a_{11}$, $\partial(a_i) = 0$ for $i \in \{3,5,8,10,11\}$. Denote by $\widehat{K}$ the 'Legendrian mirror' of K — the image of K under the map $(q,p,u) \mapsto (-q,p,-u)$. The Legendrian knots $K, \widehat{K}$ have the same classical invariants. However, they are not Legendrian isotopic, and it is possible to distinguish them by means of their DGAs. There exist homology classes ξ_+, ξ_- in the graded homology ring $H(A,\partial)$ such that $\deg(\xi_+) = 1$, $\deg(\xi_-) = -1$, and $\xi_+\xi_- = 1$ (choose $\xi_+ = [a_{10}]$, $\xi_- = [a_{11}]$). It follows from the definitions that the DGA for $\widehat{K}$ is obtained from (A,∂) by applying the anti-automorphism reversing the order of gener-

ators in all monomials. Thus, if K and $\widehat{K}$ are Legendrian isotopic then the graded homology ring $H(A,\partial)$ is anti-isomorphic to itself, and there exist $\xi'_+, \xi'_- \in H(A,\partial)$ such that $\deg(\xi'_+) = 1$, $\deg(\xi'_-) = -1$, $\xi'_- \xi'_+ = 1$. But one can check that such classes do not exist (see [14, 15] for details) and hence K and $\widehat{K}$ are not Legendrian isotopic. Note that 'first order invariants' such as Poincaré polynomials are useless in distinguishing Legendrian mirror knots.

3. Admissible decompositions of fronts

3.1. Definitions

In this section, we present the invariants of Legendrian knots constructed in [2]. These invariants are defined in terms of the front projection.

Given a σ-generic oriented Legendrian knot L, denote by $C(L)$ the set of its points corresponding to cusps of $\sigma(L)$. The Maslov index $\mu\colon L \setminus C(L) \to \Gamma = \mathbb{Z}/m(L)\mathbb{Z}$ is a locally constant function, uniquely defined up to an additive constant by the following rule: the value of μ jumps at points of $C(L)$ by ± 1 as shown in Figure 4. We call a crossing of $\Sigma = \sigma(L)$ Maslov if μ takes the same value on both its branches.

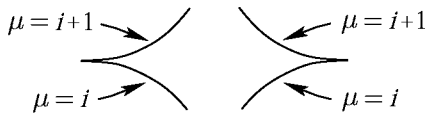


Figure 4: Jumps of the Maslov index near cusps

Assume that $\Sigma = \sigma(L)$ is a union of closed curves $X_1, \dots, X_n$ that have finitely many self-intersections and meet each other at finitely many points. Then we call the unordered collection $\{X_1, \dots, X_n\}$ a decomposition of Σ. A decomposition $\{X_1, \dots, X_n\}$ is called admissible if it satisfies certain conditions, which we are going to define. The first two are as follows:

(1) Each curve X_i bounds a topologically embedded disk: $X_i = \partial B_i$.

(2) For each $i \in \{1, \dots, n\}$, $q \in \mathbb{R}$, the set $B_i(q) = \{\, u \in \mathbb{R} \mid (q,u) \in B_i \,\}$ is either a segment, or consists of a single point u such that (q,u) is a cusp of Σ, or is empty.

Conditions (1) and (2) imply that each curve X_i has exactly two cusps (and hence the number of curves is half the number of cusps). Each X_i is divided by cusps into two pieces, on which the coordinate q is a monotone function. Near a crossing $x \in X_i \cap X_j$, the decomposition of Σ may look in one of the three ways represented in Figure 5. Conditions (1) and (2), in particular, rule out the decomposition shown in Figure 5a. We call the crossing point x switching if X_i and X_j are not smooth near x (Figure 5b), and non-switching otherwise (Figure 5c).

(3) If $(q_0, u) \in X_i \cap X_j$ $(i \neq j)$ is switching then for each $q \neq q_0$ sufficiently close to q_0 the set $B_i(q) \cap B_j(q)$ either coincides with $B_i(q)$ or $B_j(q)$, or is empty.

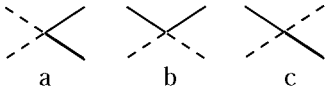


Figure 5: Local decompositions

(4) Every switching crossing is Maslov.

We call a decomposition admissible if it satisfies Conditions (1)-(3), and graded admissible if it also satisfies Condition (4). Denote by $\mathrm{Adm}(\Sigma)$ (resp. $\mathrm{Adm}_+(\Sigma)$) the set of admissible (resp. graded admissible) decompositions of Σ. Given $D \in \mathrm{Adm}(\Sigma)$, denote by $\mathrm{Sw}(D)$ the set of its switching points. Define $\theta(D) = \#(D)) - \#(\mathrm{Sw}(D))$.

Theorem 3.1. *If σ-generic Legendrian knots $L, L' \subset \mathbb{R}^3$ are Legendrian isotopic then there exists a one-to-one mapping $g\colon \mathrm{Adm}(\sigma(L)) \to \mathrm{Adm}(\sigma(L'))$ such that $g(\mathrm{Adm}_+(\sigma(L))) = \mathrm{Adm}_+(\sigma(L'))$ and $\theta(g(D)) = \theta(D)$ for each $D \in \mathrm{Adm}(\sigma(L))$. In particular, the numbers $\#(\mathrm{Adm}(\sigma(L)))$ and $\#(\mathrm{Adm}_+(\sigma(L)))$ are invariants of Legendrian isotopy.*

3.2. Remarks

1. Decompositions of fronts were first considered by Eliashberg in [4] (only Conditions (1) and (2) were involved).

2. No examples are known where the total number of admissible decompositions $\#(\mathrm{Adm}(\Sigma))$ is different for two Legendrian knots with coinciding classical invariants.

3. The proof of Theorem 3.1 goes as follows: we connect L with L' by a generic path in the space of Legendrian knots and define a canonical way to extend admissible decompositions through the points where the front is not σ-generic. The mapping g depends on the choice of the path: a loop in the space of Legendrian knots lifts to an automorphism of $\mathrm{Adm}(\sigma(L))$ which can be non-trivial even when the loop is contractible. The meaning of this phenomenon is not clear.

4. It would be interesting to understand the relation between admissible decompositions and DGAs of Legendrian knots. The first result in this direction is that if $\mathrm{Adm}_+(\sigma(L))$ is nonempty then the set $I(L)$ defined in the previous section is also nonempty [12].

3.3. Examples

Note that every admissible decomposition D of a front Σ is uniquely defined by its set of switching points. Indeed, denote by $X(\Sigma)$ the set of crossings of Σ, then each subset $E \subset X(\Sigma)$ defines a decomposition $D(E)$ of Σ which near $x \in X(\Sigma)$ has the form shown in Figure 5b if $x \in E$, and the form shown in Figure 5c otherwise. Clearly, if $E = \mathrm{Sw}(D)$ then $D = D(E)$.

The Legendrian knots represented by the fronts Σ, Σ' in Figure 6 are respectively Legendrian isotopic to the Legendrian knots L, L' defined in Figure 2. We are

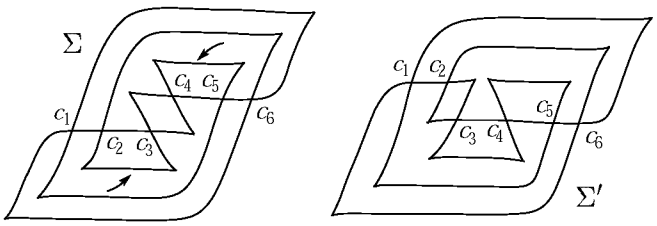


Figure 6: Fronts of two Legendrian 5_2 knots

going to show that $\#(\mathrm{Adm}_+(\Sigma)) = 1$, $\#(\mathrm{Adm}_+(\Sigma')) = 2$, and hence Theorem 1.1 is a consequence of Theorem 3.1. Assume that $D \in \mathrm{Adm}(\Sigma)$. Consider the curve $X_1 \in D$ containing the piece of Σ indicated by the lower arrow. Being applied to X_1, Conditions (1) and (2) imply that $c_2, c_3 \in \mathrm{Sw}(D)$. Similarly, looking at the curve $X_2 \in D$ containing the piece of Σ indicated by the upper arrow, we conclude that $c_4, c_5 \in \mathrm{Sw}(D)$. If one of the crossings c_1, c_6 is switching, so is the other. Then either $\mathrm{Sw}(D) = \{c_2, c_3, c_4, c_5\}$ or $\mathrm{Sw}(D) = \{c_1, c_2, c_3, c_4, c_5, c_6\}$. It is not hard to check that both decompositions are admissible but only the first one is graded. Thus $\#(\mathrm{Adm}_+(\Sigma)) = 1$. Arguing similarly, one can find that $\#(\mathrm{Adm}(\Sigma')) = 2$, where the admissible decompositions D_1, D_2 are defined by $\mathrm{Sw}(D_1) = \{c_2, c_3, c_4, c_5\}$, $\mathrm{Sw}(D_2) = \{c_1, c_2, c_3, c_4, c_5, c_6\}$, and are both graded.

4. Instability of invariants

There are two stabilizing operations, S_- and S_+, on Legendrian isotopy classes of oriented Legendrian knots, defined as follows. Given an oriented Legendrian knot L, we perform one of the operations shown in Figure 7 in a small neighbourhood of a point on L. One can check that, up to Legendrian isotopy, the resulting Legendrian knot $S_\pm(L)$ does not depend on the choices involved, and the operations S_-, S_+ commute. An important observation is that two Legendrian knots L, L' have the same classical invariants if and only if they are stable Legendrian isotopic in the sense that there exist $n_-, n_+ \in \mathbb{N}_0$ such that $S_-^{n_-}(S_+^{n_+}(L))$ is Legendrian isotopic to $S_-^{n_-}(S_+^{n_+}(L'))$ [13].

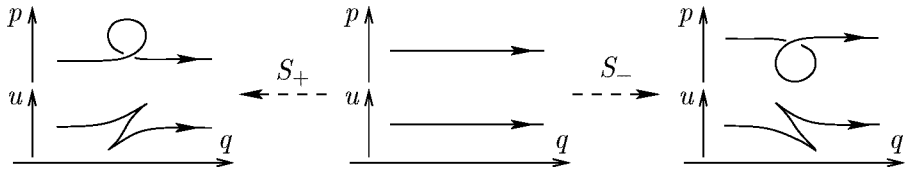


Figure 7: Stabilizations

Thus the invariants constructed in Sections 2 and 3 cannot be stable. In fact, they fail already after the first stabilization. The homology ring H of the DGA corresponding to $S_\pm(L)$ vanish, and the set $I(S_\pm(L))$ is empty. This can be easily

derived from the fact that the DGA of $S_{\pm}(L)$ can be obtained from the DGA of L by adding a new generator a such that $\partial(a) = 1$. The front of $S_{\pm}(L)$ has no admissible decompositions because Conditions (1) and (2) cannot hold for the curve X_i containing the newly created cusps.

Studying Legendrian realizations of non-prime knots, Etnyre and Honda constructed, for each m, examples of Legendrian knots that have the same classical invariants but are not Legendrian isotopic even after m stabilizations [9]. Their proof uses the classification of Legendrian torus knots given in [8]. It is an open problem to find invariants distinguishing those knots, or any pair of stabilized knots with the same classical invariants.

References

[1] Yu. V. Chekanov, Differential algebra of Legendrian links, to appear in *Inventiones Mathematicae.*
[2] Yu. V. Chekanov & P. E. Pushkar, in preparation.
[3] V. Colin, E. Giroux & K. Honda, On the coarse classification of tight contact structures, Preprint, 2002.
[4] Ya. Eliashberg, A theorem on the structure of wave fronts and its application in symplectic topology, *Funct. Anal. Appl.*, 21 (1987), 227–232.
[5] Ya. Eliashberg, Legendrian and transversal knots in tight contact 3-manifolds, In: *Topological methods in modern mathematics (Stony Brook, NY, 1991)*, Publish or Perish, 1993, 171–193.
[6] Ya. Eliashberg & M. Fraser, Classification of topologically trivial Legendrian knots, In: *Geometry, topology, and dynamics (Montreal, PQ, 1995)*, CRM Proc. Lecture Notes, 15, AMS, Providence, 1998, 17–51.
[7] Ya. Eliashberg, A. Givental, & H. Hofer, An introduction to symplectic field theory, *Geom. Funct. Anal.* (2000), Special Volume, Part II, 560–673.
[8] J. Etnyre & K. Honda, Knots and contact geometry I: torus knots and the figure eight knot, Preprint, 2000, math.GT/0006112.
[9] J. Etnyre & K. Honda, Knots and Contact Geometry II: Connected Sums, Preprint, 2002, math.GT/0205310.
[10] J. Etnyre, L. Ng, & J Sabloff, Invariants of Legendrian knots and coherent orientations, Preprint, 2001, math.GT/0101145.
[11] A. Floer, Symplectic fixed points and holomorphic spheres, *Comm. Math. Phys.*, 120, 1989, 575–611.
[12] D. Fuchs, Private communication, 2001.
[13] D. Fuchs & S. Tabachnikov, Invariants of Legendrian and transverse knots in the standard contact space, *Topology*, 36 (1997), 1025–1053.
[14] L. Ng, Legendrian mirrors and Legendrian isotopy, Preprint, 2000, math.GT/0008210.
[15] L. Ng, Computable Legendrian invariants, Preprint, 2001, math.GT/0011265.

ICM 2002 · Vol. II · 395–403

Finite Dimensional Approximations in Geometry

M. Furuta*

Abstract

In low dimensional topology, we have some invariants defined by using solutions of some nonlinear elliptic operators. The invariants could be understood as Euler class or degree in the ordinary cohomology, in infinite dimensional setting. Instead of looking at the solutions, if we can regard some kind of homotopy class of the operator itself as an invariant, then the refined version of the invariant is understood as Euler class or degree in cohomotopy theory. This idea can be carried out for the Seiberg-Witten equation on 4-dimensional manifolds and we have some applications to 4-dimensional topology.

2000 Mathematics Subject Classification: 57R57.
Keywords and Phrases: Seiberg-Witten, 4-manifold, Finite dimensional approximation.

1. Introduction

The purpose of this paper is to review the recent developments in a formal framework to extract topological information from nonlinear elliptic operators.

We also explain some applications of the idea to 4-dimensional topology by using the Seiberg-Witten theory.

A prototype is the notion of index for linear elliptic operators. In this introduction we explain this linear case. Later we mainly explain the Seiberg-Witten case.

Let $D : \Gamma(E^0) \to \Gamma(E^1)$ be an linear elliptic operator on a close manifold X. The index $ind\, D$ is defined to be

$$ind\, D = \dim \mathrm{K}er D - \dim \mathrm{C}oker D.$$

We can extend this definition as follows. Take any decomposition $D = L \oplus L' : V^0 \oplus W^0 \to V^1 \oplus W^1$. such that $L : V^0 \to V^1$ is a linear map between two finite

*Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1, Komaba, Meguro-ku, Tokyo 153-8914, Japan. E-mail: furuta@ms.u-tokyo.ac.jp

dimensional vector spaces and that $L' : W^0 \to W^1$ is an isomorphism between infinite dimensional vector spaces. Then we have

$$ind\, D = \dim V^0 - \dim V^1.$$

It is easy to check that the right-hand-side is independent of the choice of the decomposition. For example we have decomposition satisfying $V^0 = Ker D$, $V^1 \cong Coker D$, $L = 0$, which gives the former definition of the index.

An important property of $ind\, D$ is its invariance under continuous variation of D. This property is closely related to the above well-definedness.

Another way to understand this property is to consider the whole space of Fredholm maps. Then the given map D sits in the space and the $ind\, D$ is nothing but the label of the connected component containing D.

In other words, there are presumably three possible attitudes:

1. The essential data is "supported" on $Ker D$ and $Coker D$.
2. It is convenient to look at "some" finite dimensional approximation. $L : V^0 \to V^1$.
3. The essential data is the whole map $D : \Gamma(E^0) \to \Gamma(E^1)$.

When one considers a family of elliptic operators and tries to define the index of the family, it is not enough to look at their kernels and cokernels.

It is tempting to regard the finite dimensional approximation as a topological version of the notion of "low energy effective theory" in physics. In this story, the whole map D would be regarded as a given original theory.

In this paper we explain a nonlinear version of the notion of index which is formulated by using finite dimensional approximations.

2. Non-linear cases

While every elliptic operator on a closed manifold has its index as topological invariant, it is quite rare that a nonlinear elliptic operator gives some topological invariant.

We have three examples of this type of invariants: the Donaldson invariant, the Gromov-Ruan-Witten invariant and the Seiberg-Witten invariant. Moreover, the Casson invariant is regarded a variant of the Donaldson invariant. Some other finite type invariants for 3-manifolds are also supposed to be related to these kinds of invariant [37].

Even for these cases, however, it is not obvious how to proceed to obtain nonlinear version of index in full generality.

Let us first give several examples of finite dimensional approximations.

1. C. Conley and E. Zehnder solved the Arnold conjecture for torus by reducing a certain variational problem to a finite dimensional Morse theory [10].
2. Casson's definition of the Casson invariant. Taubes gave an interpretation of the Casson invariant via gauge theory [41]. In other words, Casson's construction gave a finite dimensional approximation of the gauge theoretical setting

by Taubes. (The statement of the Atiyah-Floer conjecture could be regarded as a *partial* finite dimensional approximation along fibers.)

3. Seiberg-Witten equation. The moduli space of Seiberg-Witten equation is known to be compact for closed 4-manifolds. This enables us to globalize the Kuranishi construction to obtain finite dimensional approximations [18], [3].
4. Seiberg-Witten-Floer theory C. Manolescu and P.B. Kronheimer defined Floer homotopy type for Seiberg-Witten theory, which is formulated as spectrum [32] [27].
5. Kontsevish explained an idea to define invariants of 3-manifolds by using configuration spaces. This idea was realized by Fukaya [13], Bott-Cattaneo [5] [6], and Kuperberg-Thurston [28]. Formally the configuration spaces appear as finite approximations of certain path spaces.

3. Kuranishi construction

While the index is regarded as the infinitesimal information of a nonlinear elliptic operator, its local information is given by the Kuranishi map, which has been used to describe local structure in various moduli problems [29].

A few years ago the Arnold conjecture was solved in a fairly general setting and the Gromov-Ruan-Witten invariant was defined for general symplectic manifolds. These works were done by several groups independently [14], [31], [34], [38]. A key of their arguments was to construct virtual moduli cycle over $\mathbb{Q}$.

In their case, the point is to glue local structure to obtain some global data to define invariants. Since their invariants are defined by evaluating cohomology classes, it was enough to have the virtual moduli cycle.

4. Global approximation

The notion of Fukaya-Ono's Kuranishi structure or Ruan's virtual neighborhood is defined as equivalence class of collections of maps, which define the moduli space. The collection of maps is necessary because the moduli space as topological space is not enough to recover the nature of the singularity on it.

The data depend on the choice of various choice of auxiliary data. When we change the data, the change of the moduli space is supposed to be given by a cobordism, even with the extra structure we have to look at.

Suppose we would like to regard this structure itself as an invariant. Then we have to identify the place where the invariant lives. Since cobordism classes are identified by Pontrjagin-Thom construction, what we need would be a certain stable version of Pontrjagin-Thom construction.

In the case of symplectic geometry or Donaldson's theory, this construction has not been done. A main problem seems to describe a finite dimensional approximation of the ambient space where the compactification of the moduli space lies. (The same problem occurs for Kotschick-Morgan conjecture.) Since the compactification is fairly complicated, it is not straightforward to identify the finite dimensional approximation.

However in the Seiberg-Witten case, the moduli spaces are known to be compact for closed 4-manifolds and it is not necessary to take any further compactifications.

Let us briefly recall the Seiberg-Witten equation for a closed Spin^c manifold X. For simplicity we assume $b_1(X) = 0$. Let $W = W^0 \oplus W^1$ be the spinor bundle and $\mathcal{A}$ be the space of connections on $\det W^0 \cong \det W^1$. Then the Seiberg-Witten equation is given by a map

$$\Gamma(W^0) \times \mathcal{A} \to \Gamma(W^1) \times \Gamma(\wedge^+),$$

where $\Gamma(\wedge^+)$ is the self-dual 2-forms for a fixed Riemannian metric. This is an $U(1)$-equivariant map. The inverse image of 0 divided by S^1 is the moduli space, which is known to be compact.

A finite approximation of the above map is defined by global version of the Kuranishi construction. The approximation is a *proper* $U(1)$-equivariant map

$$\mathbb{C}^{a_0} \oplus \mathbb{R}^{d_0} \to \mathbb{C}^{a_1} \oplus \mathbb{R}^{d_1}$$

for some natural numbers c_0, c_1, d_0 and d_1. The differences $c_0 - c_1$ and $d_0 - d_1$ depends only on the topology of X and its spin^c-structure.

The invariant we have is the stable homotopy class of the above $U(1)$-equivariant proper map, or equivalently, the $U(1)$-equivariant map from the sphere $S(\mathbb{C}^{a_0} \oplus \mathbb{R}^{d_0})$ to the sphere $S(\mathbb{C}^{a_1} \oplus \mathbb{R}^{d_1})$.

S. Bauer and the author pointed out that the invariant constructed above is a refinement of the usual Seiberg-Witten invariant [3].

5. 4-dimensional topology and Seiberg-Witten theory

We explain some applications of the finite dimensional approximation to 4-dimensional topology.

(1) Bauer's connected sum formula [2]

Suppose X is the connected sum of X_0 and X_1. If the neck of the connected sum is long enough, it is known that the moduli space of the solution of the Seiberg-Witten equation (or anti-self-dual equation) for X is identified with the product of the moduli spaces for X_0 and X_1. When $X_1 = \overline{CP^2}$, then this gives the blowing-up formula. When $b^+(X_0), b^-(X_1) \geq 1$, this gives vanishing of the Seiberg-Witten (or the Donaldson) invariant of $X = X_0 \# X_1$. Bauer essentially showed that the product formula holds true for the virtual neighborhood of the moduli spaces, if we use Ruan's terminology. In the language of stable maps between spheres, "product" becomes "join". In particular Bauer's formula gives the blowing-up formula for the refined invariant. When $b^+(X_0), b^-(X_1) \geq 1$, the join is torsion. It is, however, not necessary zero. In this way Bauer gave many new examples of 4-manifolds which are homeomorphic but not diffeomorphic to each other.

Ishida-Lebrun [24] [25] obtained some applications of the connected sum formula to Riemannian geometry.

(2) Intersection form of spin 4-manifolds

When 4-manifold is spin, we have certain extra symmetry, and the place where the invariant lives is a set of $Pin(2)$-equivariant stable maps [18].

When X is a closed spin 4-manifold with $b_1(X) = 0$, the Seiberg-Witten map for the spin structure is a $Pin(2)$-equivariant map formally given by

$$\mathbb{H}^\infty \oplus \tilde{\mathbb{R}}^\infty \to \mathbb{H}^\infty \oplus \tilde{\mathbb{R}}^\infty,$$

where $\tilde{\mathbb{R}}$ is the non-trivial 1-dimensional real representation space of $Pin(2)$. and $\mathbb{H}$ is the 4-dimentional real irreducible representation space of $Pin(2)$. Let $\mathbb{Z}/4$ be the subgroup of $Pin(2)$ generated by an element in $Pin(2) \setminus U(1)$. The differences of the power ∞'s are given by the index of some elliptic operators.

A finite dimensional approximation is given by a $Pin(2)$-equivariant proper map

$$\mathbb{H}^{c_0} \oplus \tilde{\mathbb{R}}^{d_0} \to \mathbb{H}^{c_1} \oplus \tilde{\mathbb{R}}^{d_1},$$

for some c_0, c_1, d_0, d_1 satisfying

$$c_0 - c_1 = -\frac{sign\,(X)}{16},\ d_0 - d_1 = b^+(X).$$

This existence implies some inequality between the signature and the second Betti number.

To obtain the inequality explicitly we can use the following results.

Theorem *Suppose $k > 0$ and $k \equiv a \bmod 4$ for $a = 0, 1, 2,$ or 3. Then there does not exist a G-equivariant continuous map from $S(\mathbb{H}^{k+x} \oplus \tilde{\mathbb{R}}^y)$ to $S(\mathbb{H}^x \oplus \tilde{\mathbb{R}}^{2k+a'-1+y})$. for the following G and a'.*

1. (B. Schmidt [39] see also [40] [11] [33]) *$G = \mathbb{Z}/4$ and $a' = a$ for $a = 1, 2, 3$.*
2. (F - Y.Kametani [20]) *$G = Pin(2)$ and $a' = 3$ for $a = 0$.*

From the above non-existence results, we have the following inequality, which is a partial result towards the 11/8-conjecture $b^+ \geq 3|sign\,(X)/16|$.

Theorem *Let X be a closed spin 4-manifold with $sign\,(X) = -16k < 0$. If $k \equiv a \bmod 4$ for $a = 0, 1, 2$ or 3, then we have $b^+ \geq 2k+b$, where $a' = a$ if $a = 1, 2, 3$ and $a' = 3$ if $a = 0$.*

Equivariant version and V-manifold version can be formulated similarly [7], [12], [16], [1]. There are some applications of these extended versions:

1. C. Bohr [4] and R. Lee - T.-J. Li [30] investigated the intersection forms of closed even 4-manifolds which are not spin.
2. Y. Fukumoto, M. Ue and the author [16] [15] [17] [42], and N. Saveliev [36] investigated homology cobordims groups of homology 3-spheres.

When $b_1 > 0$, we can construct another closed spin 4-manifold with $b_1 = 0$ without changing the intersection form. It implies that we can assume $b_1 = 0$ to obtain restriction on the intersection form. However when the intersection form on $H^1(X)$ is non-trivial, we may have a stronger restriction. Y. Kametani, H. Matsue, N. Minami and the author found that such a phenomenon actually occurs if there are $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in H^1(X, \mathbb{Z})$ such that $\langle \prod \alpha_i, [X] \rangle$ is odd [22].

6. Seiberg-Witten-Floer homotopy type

Recently C. Manolescu and P. B. Kronheimer extends the above formulation for closed 4-manifolds to the relative version [32], [27]. Let us explain their theory briefly.

We mentioned that Conley-Zehnder used a finite dimensional approximation of a Morse function on an infinite dimensional space to approach the Arnold conjecture for torus. Following this line, Conley exteded the notion of Morse index and defined the Conley index for compact isolated set [9]. The Conley index is not a number, but a homotopy type of spaces. Floer extracted some information from the Conley index just by looking at some finite dimensional skeleton of the Conley index under some assumption. Floer's formulation has the advantage that the Floer homology is defined even when the Conley index is not rigorously defined.

On the other hand R. L. Cohen, J. D. S. Jones and G. B. Segal tried to define certain stable homotopy type directly which should be an extended version of the Conley index [8]. They called it the Floer homotopy type. At that time the Floer homology was defined only for the Donaldson theory and the Gromov-Ruan-Witten theory. In these theories the moduli spaces are non-compact in general. This cause a serious difficulty to carry out their program.

In the Seiberg-Witten theory, we have a strong compactness for the moduli spaces. Manolescu and Kronheimer succeeded to construct the Floer homotopy type as spectra for the Seiberg-Witten theory by using this compactness.

They also defined relative invariant for 4-manifolds with boundary is also defined and it extends the invariant in [3].

7. Concluding remarks

The idea of finite dimensional approximation is closely related to the notion of "low energy effective theory" in physics. Actually the approximation should be regarded just as a part of the vast notion which we can deal with rigorously or mathematically.

Since Witten's realization of Donaldson theory as a TQFT, the formal relation between mathematically regorous definition of invariants and their formal path integral expressions has suggested many things. For instance, the well-definedness of the Donaldson invariant is based on the fact that the formal dimension of the moduli space increases when the instanton number goes up. This fact seems equivalent to the other fact that the pure Yang-Mills theory is asymptitotically free (for N=2 SUSY theory) and its renormalized theory does exists.

In the case of the finite dimensional approximations of Seiberg-Witten theory, the suspension maps give relations between many choices of approximations. If we use some generalized cohomology theories to detect our invariants, the suspension maps induces the Thom isomorphisms, or integrations along fibers. If we compare this setting with physics, the family of integrations look quite similar to the renormalization group. It seems the Thom classes which play the role of vacua. In this sense, one could say that the family of finite approximations is a topological version

of the renormlization group. This topological setting is very limitted. It, however, has one advantage: Usually the path integral expression is supposed to take values in real or complex numbers. On the other hand our invariants could take values in torsions.

Let us conclude this survey by giving three open problems.

1. What is the correct formulation of the geography of spin 4-manifolds with $b_1 > 0$? (If the intersection on H^1 is complicated enough, then $sign\,(X)$ would have stronger restriction.)
2. When an oriented closed 3-manifold is a link of isolated algebraic singular point, construct a canonical Galois group action on some completion of the Floer homotopy type of Kronheimer-Manolescu. (This problem was suggested by a hand-written manuscript by D. Johnson in which Casson-type invariants were defined.)
3. The Seiberg-Witten map is quadratic. Extract non-topological information from this structure. (Is it possible to approach the 11/8-conjecture from this point of view?)

References

[1] D. J. Acosta, A Furuta-like inequality for spin orbifolds and the minimal genus problem, Topology Appl. 114 (2001) 91–106.

[2] S. Bauer, A stable cohomotopy refinement of Seiberg-Witten invariants: II, math.DG/0204267

[3] S. Bauer and M. Furuta, A stable cohomotopy refinement of Seiberg-Witten invariants: I, math.DG/0204340

[4] C. Bohr, On the signatures of even 4-manifolds, *Math. Proc. Cambridge Philos. Soc.* **132** (2002), 453–469.

[5] R. Bott & A. S. Cattaneo, Integral invariants of 3-manifolds. *J. Differential Geom.* **48** (1998), no. 1, 91–133.

[6] R. Bott & A. S. Cattaneo, Integral invariants of 3-manifolds. II, *J. Differential Geom.* **53** (1999), no. 1, 1–13.

[7] J. Bryan, Seiberg-Witten theory and $Z/2^p$ actions on spin 4-manifolds, *Math. Res. Lett.* **5** (1998) 165–183.

[8] R. L. Cohen, J. D. S. Jones & G. B. Segal, Floer's infinite dimensional Morse theory and homotopy theory, The Floer memorial Volume, Birkhauser (1995).

[9] C. Conley Isolated invariant sets and the Morse index, Amer. Math. Soc., Provindence, 1978.

[10] C. Conley & E. Zehnder, The Birkhoff-Lewis fixed point theorem and a conjecture of V. I. Arnold. *Invent. Math.* **73** (1983), 33–49.

[11] M. C. Crabb, Periodicity in $\mathbb{Z}/4$-equivariant stable homotopy theory, *Cont. Math.* **96**, (1989) 109–124.

[12] F. Fang, Smooth group actions on 4-manifolds and Seiberg-Witten invariants, *Internat. J. Math.* **9** (1998), 957–973.

[13] K. Fukaya, Morse homotopy and Chern-Simons perturbation theory, *Comm. Math. Phys.* **181** (1996), no. 1, 37–90.

[14] K. Fukaya & K. Ono, Arnold conjecture and Gromov-Witten invariant. *Topology* **38** (1999), no. 5, 933–1048.
[15] Y. Fukumoto, On an invariant of plumbed homology 3-spheres,*J. Math. Kyoto. Univ.* 40, No. 2, (2000) 379–389.
[16] Y. Fukumoto & M. Furuta, Homology 3-spheres bounding acyclic 4-manifolds, *Math. Res. Lett.* **7**, (2000), 757–766.
[17] Y. Fukumoto, M. Furuta & M. Ue, W invariants and the Neumann-Siebenmann invariants for Seifert homology 3-spheres, *Topology Appl.* **116** (2001), no. 3, 333–369. .
[18] M. Furuta, Monopole equation and the 11/8-conjecture, *Math. Res. Lett.* **8** (2001), no. 3, 279–291.
[19] M. Furuta & Y. Kametani, The Seiberg-Witten equations and equivariant e-invariants, preprint (UTMS 2001-10, University of Tokyo).
[20] M. Furuta & Y. Kametani, Equivariant maps and KO^*-degree, preprint.
[21] M. Furuta, Y. Kametani & H. Matsue, Spin 4-manifolds with signature=-32, *Math. Res. Letters* **8**, (2001) 293–301.
[22] M. Furuta, Y. Kametani, H. Matsue & N. Minami, Stable-homotopy Seiberg-Witten invariants and Pin bordisms, preprint (UTMS 2000-46, University of Tokyo).
[23] M. Furuta, Y. Kametani & N. Minami, Stable-homotopy Seiberg-Witten invariants for rational cohomology $K3\#K3$'s, *J. Math. Sci. Univ. Tokyo.* **8** (2001) 157–176.
[24] M. Ishida & C. LeBrun, Spin Manifolds, Einstein Metrics, and Differential Topology, math.DG/0107111
[25] M. Ishida & C. LeBrun, Curvature, Connected Sums, and Seiberg-Witten Theory, math.DG/0111228
[26] M. Kontsevish, Feynman diagrams and low-dimensional topology, First European Congress of Mathematics, Vol. II (Paris, 1992), 97–121, Progr. Math., 120, Birkhauser, Basel, 1994.
[27] P. B. Kronheimer & C. Manolescu, Floer pro-spectra from the Seiberg-Witten equations, math.GT/0203243
[28] G. Kuperberg & D. Thurston, Perturbative 3-manifold invariants by cut-and-paste topology, math.GT/9912167
[29] M. Kuranishi, New proof for the existence of local free complete families of complex structures, Conference on Complex Analysis. Springer, Minneapolis, 1964.
[30] R. Lee & T.-J. Li, Intersection forms of non-spin four manifolds, *Math. Ann.* 319 (2001), no. 2, 311–318.
[31] G. Liu & G. Tian, Floer homology and Arnold conjecture, *J. Differential Geom.* **49** (1998), no. 1, 1–74.
[32] C. Manolescu, Seiberg-Witten-Floer stable homotopy type of three-manifolds with $b_1 = 0$, math.DG/0104024
[33] N. Minami, The G-join theorem - an unbased G-Freudenthal theorem, preprint.
[34] Y. Ruan, Virtual neighborhoods and pseudo-holomorphic curves, Proceedings of 6th Gokova Geometry-Topology Conference. *Turkish J. Math.* **23** (1999),

no. 1, 161–231.
[35] Y. Ruan, Virtual neighborhoods and the monopole equations, Topics in symplectic 4-manifolds (Irvine, CA, 1996), 101–116, First Int. Press Lect. Ser., I, Internat. Press, Cambridge, MA, 1998.
[36] N. Saveliev, Fukumoto-Furuta invariants of plumbed homology 3-spheres, preprint, (2000).
[37] N. Seiberg & E. Witten, Gauge Dynamics And Compactification To Three Dimensions, hep-th/9607163
[38] B. Siebert, Gromov-Witten invariants of general symplectic manifolds, dg-ga/9608005
[39] B. Schmidt, Ein Kriterium für die Esistenz äquivarianter Abbildungen zwischen reellen Darstellungssphären der Gruppe $Pin(2)$, Diplomarbeit Universtät Bielefeld, 1997.
[40] S. Stolz, The level of real projective spaces, *Comment. Math. Helvetici* **64** (1989), 661–674.
[41] C. H. Taubes, Casson's invariant and gauge theory, *J. Differential Geom.* **31** (1990), no. 2, 547–599.
[42] M. Ue, On the intersection forms of spin 4-manifolds bounded by spherical 3-manifolds, Algebraic and Geometric Topology 1 (2001), 549–578.

ICM 2002 · Vol. II · 405–414

Géométrie de Contact: de la Dimension Trois vers les Dimensions Supérieures

Emmanuel Giroux*

Résumé

On décrit ici des relations entre la géométrie globale des variétés de contact closes et celle de certaines variétés symplectiques, à savoir les variétés de Stein compactes. L'origine de ces relations est l'existence de livres ouverts adaptés aux structures de contact.

2000 Mathematics Subject Classification : 57M50, 53D35.
Mots clés : Structures de contact, Livres ouverts.

La géométrie de contact en dimension trois a connu un essor important durant la dernière décennie grâce au développement de méthodes topologiques adéquates. Dans le prolongement des travaux de D. Bennequin [Be] et de Y. Eliashberg [El1], la théorie des « surfaces convexes » [Gi1] et l'étude des rocades (bypasses) [Ho] ont mené à une classification complète des structures de contact sur quelques variétés simples et, plus récemment, à une classification grossière sur toutes les variétés closes [Co, HKM, CGH]. En fait, comme on essaiera de le montrer plus loin, les structures de contact en dimension trois sont des objets purement topologiques, un peu comme les structures symplectiques en dimension deux. En termes précis, sur toute variété close V de dimension trois, les classes d'isotopie des structures de contact se trouvent en correspondance bijective avec les classes d'isotopie et de stabilisation des livres ouverts dans V, l'opération élémentaire de stabilisation étant un plombage positif [Gi2].

En dimension supérieure, des méthodes radicalement différentes permettent de mettre en évidence une correspondance similaire [GM] et, au-delà, de faire apparaître des liens étroits entre la géométrie globale des variétés de contact closes et celle de certaines variétés symplectiques compactes. Les livres ouverts qu'on associe à une structure de contact sont en effet particuliers : leurs pages sont des variétés de Stein compactes, leur monodromie est un difféomorphisme symplectique à support dans l'intérieur et l'opération élémentaire de stabilisation qui les unifie est un plombage lagrangien positif. En outre, l'outil essentiel pour les construire est la théorie

*Unité de Mathématiques Pures et Appliquées, École Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon cedex 07, France. Mél: giroux@umpa.ens-lyon.fr

des fibrés positifs que S. Donaldson a introduite et développée en géométrie symplectique dans [Do1, Do2] et qui a été adaptée en géométrie de contact dans [IMP].

A. Structures de contact et livres ouverts

Dans ce texte, V désigne toujours une variété close et orientée. Les champs d'hyperplans tangents qu'on considère sur V sont coorientés, donc aussi orientés puisque V l'est. Un tel champ ξ est le noyau d'une forme α, appelée *équation* de ξ, unique à multiplication près par une fonction positive. On dit que ξ est une *structure de contact* si $d\alpha$ induit sur ξ en tout point une forme symplectique directe, *i.e.* si V est de dimension impaire $2n+1$ et si $\alpha \wedge (d\alpha)^n$ est en tout point un élément de volume direct pour l'orientation de V.

D'autre part, un *livre ouvert* dans V est un couple (K, θ) formé des objets suivants :

- une sous-variété close $K \subset V$ de codimension deux à fibré normal trivial ;
- une fibration $\theta : V \setminus K \to \mathbf{S}^1$ qui, dans un voisinage $K \times \mathbf{D}^2$ de $K = K \times \{0\}$, coïncide avec la coordonnée angulaire normale.

On peut aussi voir les livres ouverts autrement. Soit $\phi : F \to F$ un difféomorphisme d'une variété compacte égal à l'identité près du bord $K = \partial F$. Sa suspension, à savoir la variété compacte

$$\Sigma(F, \phi) = \big(F \times [0,1]\big)/\!\sim, \quad \text{où} \quad (p,1) \sim (\phi(p), 0),$$

est bordée par $K \times \mathbf{S}^1$ – car $\phi\,|_K = \mathrm{id}$ – et la variété close

$$\overline{\Sigma}(F, \phi) = \Sigma(F, \phi) \cup_\partial (K \times \mathbf{D}^2),$$

possède un livre ouvert évident. En outre, tout livre ouvert (K, θ) dans V identifie V à $\overline{\Sigma}(F, \phi)$, où F est une fibre de θ (un peu rétrécie) et ϕ l'application de premier retour sur F d'un flot transversal aux fibres de θ et constitué, près de K, de rotations autour de K. Le difféomorphisme ϕ, défini seulement à conjugaison et isotopie près, est la *monodromie* de (K, θ).

Toute la discussion à venir tourne autour de la définition suivante :

Définition 1 [Gi2, GM]. *Une structure de contact ξ sur V est dite* portée *par un livre ouvert (K, θ) si elle admet une équation α ayant les propriétés suivantes :*

- *α induit sur K une forme de contact ;*
- *$d\alpha$ induit sur chaque fibre F de θ une forme symplectique ;*
- *l'orientation de K définie par la forme de contact α coïncide avec son orientation comme bord de la variété symplectique $(F, d\alpha)$.*

Une telle forme α sera dite adaptée *à (K, θ).*

Exemple [GM]. Soit $f : (\mathbf{C}^n, 0) \to (\mathbf{C}, 0)$ une fonction holomorphe ayant à l'origine un point critique isolé et soit H l'hypersurface (singulière) $f^{-1}(0)$. Il existe une boule fermée lisse B autour de l'origine dans $\mathbf{C}^n$ et un feuilletage de $B \setminus \{0\}$ par des sphères strictement pseudoconvexes S_r, où $r \in]0,1]$ et $S_1 = \partial B$, tels que, pour r assez petit, les propriétés suivantes soient satisfaites :

- la sphère S_r est transversale à H, de sorte que $K = H \cap S_r$ est une sous-variété close de S_r de codimension deux et à fibré normal trivial ;
- l'application $\theta = \arg f \colon S_r \setminus K \to \mathbf{S}^1$ est une fibration qui fait de (K, θ) un livre ouvert ;
- le livre ouvert (K, θ) porte la structure de contact sur S_r définie par le champ des tangentes complexes.

Autrement dit, chaque livre ouvert donné dans la sphère par le théorème de fibration de J. Milnor porte, à isotopie près, la structure de contact standard.

B. Structures de contact et livres ouverts en dimension trois

En dimension trois, divers travaux ont depuis longtemps fait apparaître des connivences entre les structures de contact et les livres ouverts sans toutefois établir aucun lien formel. Dans [TW], W. Thurston et H. Winkelnkemper construisent des formes de contact sur toute variété close V à partir d'un livre ouvert dans V. Avec les termes de la définition 1, ils démontrent en fait que tout livre ouvert dans V porte une structure de contact. Dans [Be] d'autre part, pour transformer en théorème de géométrie de contact son résultat sur les tresses fermées, D. Bennequin met en évidence la propriété suivante : toute courbe transversale à la structure de contact standard ξ_0 dans $\mathbf{R}^3$ – structure d'équation $dz + r^2 d\theta = 0$ – est isotope, parmi les courbes transversales, à une tresse fermée c'est-à-dire une courbe transversale au livre ouvert formé par l'axe des z et la coordonnée angulaire θ. Or cette propriété vient de ce que ce livre ouvert porte ξ_0. Enfin, dans [To], I. Torisu a clairement dégagé les relations entre les livres ouverts et les configurations de théorie de Morse considérées dans [Gi1] pour étudier les structures de contact *convexes* au sens de [EG].

La première observation qui montre l'étroitesse des liens imposés par la définition 1 et découle de la stabilité des structures de contact est la suivante :

Proposition 2 [Gi2]. *Sur une variété close de dimension trois, toutes les structures de contact portées par un même livre ouvert sont isotopes.*

Quant à la question de savoir quelles structures de contact possèdent un livre ouvert porteur, la réponse est simple :

Théorème 3 [Gi2]. *Sur une variété close de dimension trois, toute structure de contact est portée par un livre ouvert.*

Cependant, comme l'illustre l'exemple des fibrations de Milnor, le livre ouvert qui porte une structure de contact donnée est loin d'être unique – même à isotopie près. Pour appréhender ce phénomène, quelques définitions sont utiles.

Soit $F \subset V$ une surface compacte à bord et $C \subset F$ un arc simple et propre. On dit qu'une surface compacte $F' \subset V$ s'obtient à partir de F par le *plombage positif* (resp. *négatif*) d'un anneau le long de C si $F' = F \cup A$ où $A \subset V$ est un anneau ayant les propriétés suivantes :

- $A \cap F$ est un voisinage régulier de C dans F ;

- A est inclus dans une boule fermée B dont l'intersection avec F est réduite à $A \cap F$ et l'enlacement des deux composantes de ∂A dans B vaut 1 (resp. -1).

Un résultat de J. Stallings affirme que, si (K, θ) est un livre ouvert dans V et si F est l'adhérence d'une fibre de θ, alors, pour toute surface F' obtenue à partir de F par le plombage d'un anneau, il existe un livre ouvert (K', θ') tel que K' soit le bord de F' et que F' soit l'adhérence d'une fibre de θ'. Dans la suite, on dira que le livre (K', θ') et l'entrelacs K' sont eux-mêmes obtenus par plombage à partir respectivement de (K, θ) et de K. En outre, on dira qu'un livre ouvert (K', θ') est une *stabilisation* d'un autre (K, θ) s'il s'obtient à partir de (K, θ) par une suite finie de plombages positifs.

Théorème 4 [Gi2]. *Dans une variété close de dimension trois, deux livres ouverts quelconques qui portent une même structure de contact ont des stabilisations isotopes.*

Les théorèmes 3 et 4 permettent de traduire nombre de questions sur les structures de contact en questions sur les livres ouverts, autrement dit sur les difféomorphismes des surfaces compactes à bord. En ce sens, ce sont les analogues des théorèmes de S. Donaldson [Do2] sur les pinceaux de Lefschetz dans les variétés symplectiques de dimension quatre. Ils admettent cependant, à la différence de ceux-ci, des démonstrations purement topologiques dont on décrit brièvement les idées ci-dessous, après avoir introduit l'outil essentiel. On supposera le lecteur familier avec certaines notions de géométrie de contact en dimension trois (structures de contact vrillées/tendues, invariant de Thurston-Bennequin des courbes legendriennes, surfaces ξ-convexes).

On appelle *cellule polyédrale* dans V l'image d'un polyèdre convexe compact euclidien par un plongement topologique. Une telle cellule possède une structure affine induite par son paramétrage et son *intérieur* est, par définition, l'image de l'intérieur « intrinsèque » du polyèdre, c'est-à-dire de son intérieur topologique dans son enveloppe affine. Une *cellulation polyédrale* de V désigne ici un recouvrement fini de V par des cellules polyédrales ayant les propriétés suivantes :

- les intérieurs des cellules forment une partition de V ;
- le bord de chaque cellule D est une union de cellules D_j et les inclusions $D_j \to D$ sont affines ;
- les cellules de dimension deux (et moins) sont lisses, *i.e.* sont les images de plongements lisses.

Les cellulations polyédrales ont cet avantage sur les triangulations d'être très faciles à subdiviser : toute subdivision d'un sous-complexe se prolonge trivialement. En outre, elles jouent un rôle clé dans la démonstration du théorème de Reidemeister-Singer donnée dans [Si], démonstration qui sert de guide pour établir le théorème 4.

Esquisse de la démonstration du théorème 3. Soit ξ une structure de contact sur V. On construit d'abord dans (V, ξ) une *cellulation de contact*, c'est-à-dire une cellulation polyédrale Δ ayant les propriétés suivantes :

1) chaque cellule de dimension 1 est un arc legendrien ;
2) chaque cellule de dimension 2 est ξ-convexe et l'invariant de Thurston-Bennequin de son bord vaut -1 ;

3) chaque cellule de dimension 3 est contenue dans le domaine d'une carte de Darboux.

On épaissit ensuite le 1-squelette L de Δ en une surface compacte $\hat{F}$ (presque) tangente à ξ le long de L et on choisit un voisinage régulier W de L assez petit pour que $F = \hat{F} \cap W$ soit une surface proprement plongée dans W. Quitte à prendre W plus petit, ξ admet une équation α vérifiant les conditions suivantes :

- $d\alpha$ induit sur F une forme d'aire ;
- α est non singulière sur $K = \partial F$ et oriente K comme le bord de $(F, d\alpha)$.

D'autre part, pour toute cellule D de dimension 2, la propriété 2) dit que le bord de $D \cap (V \setminus \operatorname{Int} W)$ intersecte K en deux points (à isotopie près). Il en résulte qu'il existe une fibration $\theta : V \setminus K \to \mathbf{S}^1$ ayant $\operatorname{Int} F$ pour fibre. Quitte à rogner W, on peut supposer que W est une union de fibres de θ sur lesquelles $d\alpha$ induit une forme d'aire. Il reste à voir que ξ est isotope, relativement à W, à une structure de contact portée par (K, θ). Le point clé est que ξ est tendue sur $W^* = V \setminus \operatorname{Int} W$ et que ∂W^* est une surface ξ-convexe dont le découpage est fourni par K. □

Étapes de la démonstration du théorème 4. Soit Δ une cellulation de contact de (V, ξ). On dira ici qu'un livre ouvert porteur (K, θ) est *associé* à Δ si, comme dans la démonstration du théorème 3, l'une des fibres de θ contient le 1-squelette de Δ et se rétracte dessus par une isotopie de contact. En imitant [Si], on montre d'abord que tout livre ouvert porteur admet une stabilisation associée à une cellulation de contact. On se ramène ainsi à considérer le cas de deux livres ouverts porteurs associés à des cellulations de contact Δ_0 et Δ_1 en position générale. D'après [Si], Δ_0 et Δ_1 possèdent une subdivision commune Δ_2 qui s'obtient, à partir de Δ_0 comme de Δ_1, par des bissections. On déforme alors Δ_2, relativement à l'union des 1-squelettes de Δ_0 et Δ_1, en une cellulation vérifiant les propriétés 1) et 3) des cellulations de contact et ayant des 2-cellules ξ-convexes. Il suffit ensuite de subdiviser le 2-squelette de Δ_2 pour obtenir une cellulation de contact Δ et on montre pour finir que le livre ouvert associé à Δ est une stabilisation de ceux associés à Δ_0 et à Δ_1. □

On discute maintenant quelques corollaires des théorèmes 3 et 4.

On rappelle d'abord qu'un théorème de M. Hilden et J. Montesinos affirme que toute variété close V de dimension trois est un revêtement à trois feuillets de la sphère $\mathbf{S}^3$ simplement ramifié au-dessus d'un entrelacs (simplement signifie que le degré local aux points de ramification dans V vaut deux). On obtient le même résultat pour les variétés de contact closes :

Corollaire 5 [Gi2]. *Toute variété de contact close de dimension trois est un revêtement à trois feuillets de la sphère de contact standard* $(\mathbf{S}^3, \xi_0)$ *simplement ramifié au-dessus d'un entrelacs transversal à* ξ_0.

Un autre corollaire concerne la dynamique des flots de Reeb. Un *flot de Reeb* sur une variété de contact est un flot qui préserve la structure de contact tout en lui étant transversal et en pointant du côté positif. Un exemple typique est le flot géodésique sur le fibré cotangent unitaire d'une variété riemannienne. Les flots de Reeb d'une structure de contact donnée ξ sont en bijection avec les équations de ξ : à toute forme α correspond l'unique champ de vecteurs ∇_α qui engendre le noyau

de $d\alpha$ et sur lequel α vaut 1. En prenant une équation de ξ adaptée à un livre ouvert porteur, on obtient :

Corollaire 6 [Gi2]. *Sur toute variété de contact close de dimension trois, il existe un flot de Reeb qui admet une section de Poincaré-Birkhoff, c'est-à-dire une surface compacte qui rencontre toutes les orbites, dont l'intérieur est transversal au flot et dont chaque composante du bord est une orbite périodique.*

En fait, il n'est pas exclu que tout flot de Reeb admette une telle section [HWZ] (ceci impliquerait la conjecture de Weinstein selon laquelle tout flot de Reeb a une orbite périodique) mais c'est là un problème de nature différente, certainement inaccessible par des méthodes topologiques.

Une question naturelle au vu des théorèmes 3 et 4 est de savoir comment lire sur la monodromie de ses livres ouverts porteurs si une structure de contact est tendue, ou remplissable en un quelconque sens. La seule réponse connue concerne les structures de contact *holomorphiquement remplissables*, c'est-à-dire réalisables comme champs des tangentes complexes au bord de variétés de Stein compactes. Le corollaire suivant précise un résultat de A. Loi et R. Piergallini :

Corollaire 7 [LP, Gi2]. *Une structure de contact sur une variété close de dimension trois est holomorphiquement remplissable si et seulement si elle est portée par un livre ouvert dont la monodromie est un produit de twists de Dehn à droite.*

Pour finir, on donne un corollaire de pure théorie des nœuds. On appelle ici *entrelacs fibré* dans V tout entrelacs orienté K pour lequel il existe une fibration $\theta\colon V \setminus K \to \mathbf{S}^1$ qui fait de $(K'\theta)$ un livre ouvert et induit sur K l'orientation prescrite. Lorsque V est une sphère d'homologie, un théorème de F. Waldhausen assure que cette fibration, si elle existe, est unique à isotopie près. Le résultat suivant répond à une question posée par J. Harer dans [Ha] :

Corollaire 8 [Gi2]. *Deux entrelacs fibrés quelconques dans une sphère d'homologie entière V s'obtiennent l'un à partir de l'autre par une suite de plombages et de « déplombages » (opérations inverses).*

Démonstration. Une trivialisation de V étant choisie, les classes d'homotopie de champs de plans tangents à V sont repérées par leur *invariant de Hopf*, à savoir l'enlacement des fibres des applications $V \to \mathbf{S}^2$ correspondantes. On considère alors un livre ouvert quelconque (K,θ) dans V et on note (K',θ') un livre ouvert obtenu à partir de (K,θ) par un plombage négatif. Les trois observations suivantes démontrent le corollaire :

- toute structure de contact ξ' portée par (K',θ') est vrillée car l'âme de l'anneau plombé est isotope à une courbe legendrienne non nouée dans (V,ξ') dont l'invariant de Thurston-Bennequin vaut $+1$;
- l'invariant de Hopf de ξ' est supérieur d'une unité à celui des structures de contact portées par (K,θ) (voir [NR]) ;
- si deux structures de contact vrillées ont le même invariant de Hopf, elles sont isotopes d'après [El1] et deux livres ouverts quelconques qui les portent ont donc des stabilisations isotopes.

Cet argument borne en outre par $h+2$ le nombre des (dé) plombages négatifs nécessaires pour passer d'un livre ouvert à un autre, où h désigne la différence

entre les invariants de Hopf correspondants. □

C. Structures de contact et livres ouverts en dimension supérieure

En dimension supérieure à trois, les livres ouverts porteurs de structures de contact ne sont pas quelconques : leurs fibres ont une structure symplectique invariante par la monodromie. Pour préciser ce point, quelques définitions sont utiles.

Soit F une variété compacte, à bord $K = \partial F$. Une forme symplectique exacte ω sur $\operatorname{Int} F$ est *convexe à l'infini* s'il existe sur $\operatorname{Int} F$ un champ de Liouville (champ de vecteurs ω-dual d'une primitive de ω) qui est transversal à toutes les hypersurfaces $K \times \{t\}$, $t \in]0,1]$, où $K \times [0,1]$ est un voisinage collier de $K = K \times \{0\}$. On dit en outre que $(\operatorname{Int} F, \omega)$ est une *variété de Weinstein* [EG] s'il existe un tel champ de Liouville qui, de plus, est le (pseudo) gradient d'une fonction de Morse $F \to \mathbf{R}$ constante et sans points critiques sur K. L'exemple typique de variété de Weinstein est l'intérieur d'une *variété de Stein compacte.* On nomme ainsi toute variété complexe compacte F qui admet une fonction strictement pluri-sous-harmonique $f\colon F \to \mathbf{R}$ constante et sans points critiques sur le bord. La 2-forme $i\partial\overline{\partial} f$ définit alors une structure symplectique. Il ressort en fait du travail de Y. Eliashberg [El2] que toute variété de Weinstein est symplectiquement difféomorphe à l'intérieur d'une telle variété de Stein compacte.

Si maintenant α est une forme de contact adaptée à un livre ouvert (K, θ), sa différentielle $d\alpha$ induit sur chaque fibre de θ une structure symplectique exacte convexe à l'infini. Celle-ci dépend du choix de α mais sa complétion [EG] est bien définie à isotopie près. Le théorème de W. Thurston et H. Winkelnkemper et la proposition 2 s'étendent alors ainsi en grande dimension :

Proposition 9 [GM]. *Soit F une variété compacte avec, sur $\operatorname{Int} F$, une forme symplectique exacte convexe à l'infini et soit $\phi\colon F \to F$ un difféomorphisme symplectique égal à l'identité près de $K = \partial F$. Il existe alors sur $\overline{\Sigma}(F, \phi)$ une structure de contact portée par le livre ouvert évident. De plus, deux structures de contact portées par un même livre ouvert et qui induisent sur ses pages des structures symplectiques ayant des complétions isotopes sont isotopes.*

Quant au théorème 3, il se généralise comme suit :

Théorème 10 [GM]. *Toute structure de contact sur une variété close V est portée par un livre ouvert dont chaque fibre est une variété de Weinstein.*

Esquisse de la démonstration. Soit ξ une structure de contact, α une équation de ξ et J une structure presque complexe sur ξ calibrée par $d\alpha\,|_{\xi}$. On note ∇_α le champ de Reeb associé à α et g la métrique riemannienne sur V qui vaut $d\alpha(.,J.)$ sur ξ et rend ∇_α unitaire et orthogonal à ξ. En termes élémentaires, le théorème principal de [IMP] montre qu'il existe des constantes $C, \eta > 0$ et des fonctions $s_k\colon V \to \mathbf{C}$, $k \geq 1$, vérifiant les conditions suivantes :

– en tout point de V,

$$|s_k(p)| \leq C, \quad |ds_k - iks_k\alpha| \leq Ck^{1/2} \quad \text{et} \quad |\overline{\partial}_\xi s_k| \leq C\,;$$

– en tout point p où $|s_k(p)| \leq \eta$,

$$|\partial_\xi s_k(p)| \geq \eta k^{1/2}\,.$$

(Ici, $\partial_\xi s_k$ et $\overline{\partial}_\xi s_k$ sont les parties respectivement J-linéaire et J-antilinéaire de $ds_k\,|_\xi$.) En termes plus parlants, les fonctions s_k sont des sections approximativement holomorphes et équitransversales du fibré $L^{\otimes k} \to V$, où L est le fibré hermitien trivial $V \times \mathbf{C} \to V$ muni de la connexion unitaire définie par la forme $-i\alpha$.

Les estimations ci-dessus entraînent d'abord que, pour $|w| \leq \eta$, l'ensemble $K_w = s_k^{-1}(w)$ est une sous-variété et que la forme α_w induite par α sur K_w est une forme de contact (voir [IMP]). En effet, α_w est non singulière pour k assez grand puisque son noyau est égal au noyau de $ds_k\,|_\xi$ et que $|\partial_\xi s_k| \geq \eta k^{1/2}$ tandis que $|\overline{\partial}_\xi s_k| \leq C$. Mieux, ces inégalités montrent que, pour k grand,le noyau de α_w est proche d'un sous-espace J-complexe de ξ si bien que $d\alpha_w$ y est non dégénérée.

L'observation suivante est que l'application $\arg s_k \colon V \setminus K \to \mathbf{S}^1$ est une fibration dont les fibres sont transversales au champ de Reeb ∇_α en tout point où $|s_k| \geq \eta$. Pour le voir, on note que l'estimation sur $ds_k - iks_k\alpha$ implique que

$$|ds_k(\nabla_\alpha) - iks_k| \leq Ck^{1/2}.$$

Ainsi, en un point p où $|s_k(p)| \geq \eta$ et pour k assez grand, $ds_k(p)\,(\nabla_\alpha)$ est proche de $iks_k(p)$, *i.e.* est non nul et presque orthogonal à $s_k(p)$. Par suite, les sous-variétés

$$s_k^{-1}(R_\theta), \quad \text{où} \quad R_\theta = \{re^{i\theta},\ r > \eta\},$$

sont transversales au champ de Reeb ∇_α.

Ces arguments montrent que le livre ouvert ($K = K_0$, $\theta = \arg s_k$), pour k assez grand, porte la structure de contact $\xi = \ker\alpha$. Il reste à vérifier que les fibres de θ sont des variétés de Weinstein. Pour simplifier, on prouve ci-dessous l'assertion analogue en géométrie symplectique. □

Proposition 11. *Soit W une variété close, ω une forme symplectique entière sur W et H_k une sous-variété symplectique de W en dualité de Poincaré avec $k\omega$ et obtenue par la construction de Donaldson* [Do1]*, à partir d'un fibré hermitien en droites L muni d'une connexion unitaire de courbure $-i\omega$. Pour k assez grand, $(W \setminus H_k, \omega)$ est une variété de Weinstein.*

Démonstration. En reprenant les arguments de [Do2], on peut supposer que H_k est le lieu d'annulation d'une section $s_k \colon V \to L^{\otimes k}$ qui vérifie, en tout point de W,

$$|\overline{\partial}_k s_k| \leq c|\partial_k s_k| \quad \text{avec} \quad c < \frac{1}{\sqrt{2}}.$$

Dans la trivialisation de $L^{\otimes k}$ donnée au-dessus de $W \setminus H_k$ par la section unitaire $u = s_k/|s_k|$, la connexion est définie par une 1-forme $-i\lambda$ où $d\lambda = k\omega$. Si on pose $s_k = \varphi u$, l'inégalité ci-dessus donne

$$|d\varphi/\varphi + J^*\lambda| < |d\varphi/\varphi - J^*\lambda|,$$

ce qui montre que $J^*\lambda$ est plus loin de $d\varphi/\varphi$ que de $-d\varphi/\varphi$. Le champ de Liouville dual de λ est alors un pseudogradient de $\log\varphi$. □

Comme en dimension trois, le livre ouvert porteur d'une structure de contact donnée n'est pas unique. On décrit dans [GM] une opération de *plombage le long d'un disque lagrangien* – dans laquelle les twists de Dehn-Seidel viennent remplacer les twists de Dehn – qui permet d'établir des analogues du théorème 4 et du corollaire 7. Ces résultats ramènent l'étude des structures de contact à celles des difféomorphismes symplectiques des variétés de Stein compactes qui sont l'identité près du bord. Ils permettent peut-être ainsi de rapprocher les travaux de Y. Eliashberg, H. Hofer et A. Givental sur la théorie symplectique des champs de ceux de, par exemple, de P. Seidel sur l'homologie de Floer et les groupes de difféomorphismes symplectiques. On peut aussi se demander si le théorème 10 cache des obstructions à l'existence d'une structure de contact sur les variétés closes. D'après [Qu], toute variété close V de dimension $2n+1$ possède un livre ouvert dont chaque fibre a le type d'homotopie d'un complexe cellulaire de dimension n. Il est probable que, si V admet un champ d'hyperplans tangents muni d'une structure presque complexe, il existe un tel livre ouvert pour lequel chaque fibre est une variété presque complexe et est donc, d'après [El2], l'intérieur d'une variété de Stein compacte. Toute la difficulté serait donc vraiment de réaliser la monodromie par un difféomorphisme symplectique... Dans cet ordre d'idée, voici un corollaire concret du théorème 10 obtenu par F. Bourgeois et qui montre, en réponse à une vieille question, que tout tore de dimension impaire possède une structure de contact :

Corollaire 12 [Bo]. *Si une variété close V admet une structure de contact, $V \times \mathbf{T}^2$ en admet une aussi.*

Démonstration. Soit ξ une structure de contact sur V, soit α une équation de ξ adaptée à un livre ouvert porteur (K,θ) et soit $N = K \times \mathbf{D}^2$ un voisinage de $K = K \times \{0\}$ dans lequel θ est la coordonnée angulaire normale. On note r la coordonnée radiale normale dans N et on pose

$$\tilde{\alpha} = \alpha + f(r)(\cos\theta\, dx_1 - sin\theta\, dx_2), \qquad (x_1, x_2) \in \mathbf{T}^2 = \mathbf{R}^2/\mathbf{Z}^2,$$

où la fonction$f(r)$ vaut r pour $r \leq r_0$, 1 pour $r \geq 2r_0$ et vérifie $f'(r) \geq 0$. Un calcul montre que, si on choisit r_0 assez petit, $\tilde{\alpha}$ est une forme de contact sur $V \times \mathbf{T}^2$. □

Références

[Be] D. Bennequin, *Entrelacements et équations de Pfaff.* Astérisque **107–108** (1983), 87–161.

[Bo] F. Bourgeois, *Odd-dimensional tori are contact manifolds.* Int. Math. Res. Notices (à paraître).

[Co] V. Colin, *Une infinité de structures de contact tendues sur les variétés toroïdales.* Comment. Math. Helv. **76** (2001), 353–372.

[CGH] V. Colin, E. Giroux et K. Honda, *Finitude homotopique et isotopique des structures de contact tendues.* En préparation.

[Do1] S. Donaldson, *Symplectic submanifolds and almost-complex geometry.* J. Diff. Geom. **44** (1996), 666–705.

[Do2] S. Donaldson, *Lefschetz pencils on symplectic manifolds.* J. Diff. Geom. **53** (1999), 205–236.

[El1] Y. ELIASHBERG, *Classification of over-twisted contact structures on 3-manifolds.* Invent. Math. **98** (1989), 623–637.

[El2] Y. ELIASHBERG, *Topological characterization of Stein manifolds of dimension* > 2. Int. J. Math. **1** (1990), 29–46.

[EG] Y. ELIASHBERG et M. GROMOV, *Convex symplectic manifolds.* Several Complex Variables and Complex Geometry (part 2), Proc. Sympos. Pure Math. **52**, Amer. Math. Soc. 1991, 135–162.

[Gi1] E. GIROUX, *Convexité en topologie de contact.* Comment. Math. Helv. **66** (1991), 637–677.

[Gi2] E. GIROUX, *Structures de contact, livres ouverts et tresses fermées.* En préparation.

[GM] E. GIROUX et J.-P. MOHSEN, *Structures de contact et fibrations symplectiques au-dessus du cercle.* En préparation.

[Ha] J. HARER, *How to construct all fibered knots and links.* Topology **21** (1982), 263–280.

[Ho] K. HONDA, *On the classification of tight contact structures I.* Geom. Topol. **4** (2000), 309–368.

[HKM] K. HONDA, W. KAZEZ et G. MATIĆ, *Convex decomposition theory.* Int. Math. Res. Notices 2002, 55–88.

[HWZ] H. HOFER, K. WYSOCKI et E. ZEHNDER, *The dynamics on three-dimensional strictly convex energy surfaces.* Ann. of Math. **148** (1998), 197–289.

[IMP] A. IBORT, D. MARTÍNEZ et F. PRESAS, *On the construction of contact submanifolds with prescribed topology.* J. Diff. Geom. **56** (2000), 235–283.

[LP] A. LOI et R. PIERGALLINI, *Compact Stein surfaces with boundary as branched covers of* B^4. Invent. Math. **143** (2001), 325–348.

[NR] W. NEUMANN et L. RUDOLPH, *Unfoldings in knot theory.* Math. Ann. **278** (1987), 409–439 – Corrigendum : Math. Ann. **282** (1988), 349–351.

[Qu] F. QUINN, *Open book decompositions and the bordism of automorphisms.* Topology **18** (1979), 55–73.

[Si] L. SIEBENMANN, *Les bissections expliquent le théorème de Reidemeister-Singer.* Prépublication 1979 (Orsay).

[To] I. TORISU, *Convex contact structures and fibered links in 3-manifolds.* Int. Math. Res. Notices **2000**, 441–454.

[TW] W. THURSTON et H. WINKELNKEMPER, *On the existence of contact forms.* Proc. Amer. Math. Soc. **52** (1975), 345–347.

ICM 2002 · Vol. II · 415–425

Algebraic K-theory and Trace Invariants

Lars Hesselholt*

(*Dedicated to Ib Madsen on his sixtieth birthday*)

Abstract

The cyclotomic trace of Bökstedt-Hsiang-Madsen, the subject of Bökstedt's lecture at the congress in Kyoto, is a map of pro-abelian groups

$$K_*(A) \xrightarrow{\mathrm{tr}} \mathrm{TR}^{\cdot}_*(A;p)$$

from Quillen's algebraic K-theory to a topological refinement of Connes' cyclic homology. Over the last decade, our understanding of the target and its relation to K-theory has been significantly advanced. This and possible future development is the topic of my lecture.

The cyclotomic trace takes values in the subset fixed by an operator F called the Frobenius. It is known that the induced map

$$K_*(A, \mathbb{Z}/p^v) \xrightarrow{\mathrm{tr}} \mathrm{TR}^{\cdot}_*(A;p,\mathbb{Z}/p^v)^{F=1}$$

is an isomorphism, for instance, if A is a regular local $\mathbb{F}_p$-algebra, or if A is a henselian discrete valuation ring of mixed characteristic $(0,p)$ with a separably closed residue field. It is possible to evaluate K-theory by means of the cyclotomic trace for a wider class of rings, but the precise connection becomes slightly more complicated to spell out.

The pro-abelian groups $\mathrm{TR}^{\cdot}_*(A;p)$ are typically very large. But they come equipped with a number of operators, and the combined algebraic structure is quite rigid. There is a universal example of this structure — the de Rham-Witt complex — which was first considered by Bloch-Deligne-Illusie in connection with Grothendieck's crystalline cohomology. In general, the canonical map

$$W.\,\Omega^q_A \to \mathrm{TR}^{\cdot}_q(A;p)$$

is an isomorphism, if $q \leq 1$, and the higher groups, too, can often be expressed in terms of the de Rham-Witt groups. This is true, for example, if A is a regular $\mathbb{F}_p$-algebra, or if A is a smooth algebra over the ring of integers in a local number field. The calculation in the latter case verifies the Lichtenbaum-Quillen conjecture for local number fields, or more generally, for henselian discrete valuation fields of geometric type.

2000 Mathematics Subject Classification: 19D45, 19D50, 19D55, 11S70, 14F30, 55P91.

*Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: larsh@math.mit.edu

1. Algebraic K-theory

The algebraic K-theory of Quillen [30], inherently, is a multiplicative theory. Trace invariants allow the study of this theory by embedding it in an additive theory. It is possible, by this approach, to evaluate the K-theory (with coefficients) of henselian discrete valuation fields of mixed characteristic. We first recall the expected value of the K-groups of a field k.

The groups $K_*(k)$ form a connected anti-commutative graded ring, there is a canonical isomorphism $\ell\colon k^* \xrightarrow{\sim} K_1(k)$, and $\ell(x) \cdot \ell(1-x) = 0$. One defines the Milnor K-groups $K_*^M(k)$ to be the universal example of this algebraic structure [29]. The canonical map $K_q^M(k) \to K_q(k)$ is an isomorphism, if $q \leq 2$. Let us now fix the attention on the K-groups with finite coefficients. (The rational K-groups, while of great interest, are of a rather different nature [11, 12].) The groups $K_*(k, \mathbb{Z}/m)$ form an anti-commutative graded $\mathbb{Z}/m$-algebra, at least if $v_2(m) \neq 1, 2$ and $v_3(m) \neq 1$. And if $\mu_m \subset k$, there is a canonical lifting

$$\begin{array}{ccc} & & K_2(k, \mathbb{Z}/m) \\ & \overset{b}{\nearrow} & \big\downarrow \beta_m \\ \mu_m & \xrightarrow{\ell} & K_1(k), \end{array}$$

which to a primitive mth root of unity ζ associates the Bott element b_ζ. Hence, in this case, there is an additional map of graded rings $S_{\mathbb{Z}/m}(\mu_m) \to K_*(k, \mathbb{Z}/m)$. The Beilinson-Licthenbaum conjectures predict that the combined map

$$K_*^M(k) \otimes_{\mathbb{Z}} S_{\mathbb{Z}/m}(\mu_m) \to K_*(k, \mathbb{Z}/m)$$

be an isomorphism of graded rings [1, 26]. The case $m = 2^v$ follows from the celebrated proof of the Milnor conjecture by Voevodsky [34]. We here consider the case of a henselian discrete valuation field of mixed characteristic $(0, p)$ with p odd and $m = p^v$ [20, 14]. The groups $K_*^M(k)/m$ typically are non-zero in only finitely many degrees. Hence, above this range, the groups $K_*(k, \mathbb{Z}/m)$ are two-periodic. All rings (resp. graded rings, resp. monoids) considered in this paper are assumed commutative (resp. anti-commutative, resp. commutative) and unital without further notice.

2. The de Rham-Witt complex

Let V be a henselian discrete valuation ring with quotient field K of characteristic zero and residue field k of odd characteristic p. (At this writing, we further require that V be of geometric type, i.e. that V be the henselian local ring at the generic point of the special fiber of a smooth scheme over a henselian discrete valuation ring $V_0 \subset V$ with *perfect* residue field.) A first example of a trace map is provided by the logarithmic derivative

$$K_*^M(K) \to \Omega^*_{(V,M)}$$

which to the symbol $\{a_1, \dots, a_q\}$ associates the form $d\log a_1 \dots d\log a_q$. The right hand side is the de Rham complex with log poles in the sense of Kato [25]: A log ring (A, M) is a ring A and a map of monoids $\alpha\colon M \to (A, \cdot)$; a log differential graded ring (E^*, M) is a differential graded ring E^* together with maps of monoids $\alpha\colon M \to (E^0, \cdot)$ and $d\log\colon M \to (E^1, +)$ such that $d \circ d\log = 0$ and such that $d\alpha(a) = \alpha(a) d\log a$ for all $a \in M$; the de Rham complex $\Omega^*_{(A,M)}$ is the universal log differential graded ring with underlying log ring (A, M). We will always consider the ring V with the *canonical* log structure

$$\alpha\colon M = V \cap K^* \hookrightarrow V.$$

(In this case, there are natural short-exact sequences

$$0 \to \Omega^q_V \to \Omega^q_{(V,M)} \to \Omega^{q-1}_k \to 0.)$$

The logarithmic derivative, however, is far from injective. It turns out that this can be rectified by incorborating the Witt vector construction which we now recall.

The ring of Witt vectors associated with a ring A is the set of "vectors"

$$W(A) = \{(a_0, a_1, \dots) \mid a_i \in A\}$$

with a new ring structure, see [19]. The ring operations are polynomial in the coordinates. The projection $W(A) \to A$, which to $(a_0, a_1, \dots)$ associates a_0, is a natural ring homomorphism with the unique natural multiplicative section

$$[\ ]\colon A \to W(A), \qquad [a] = (a, 0, 0, \dots).$$

If F is a perfect field of characteristic $p > 0$, then $W(F)$ is the unique (up to unique isomorphism) complete discrete valuation ring of mixed characteristic $(0, p)$ such that $W(F)/p \xrightarrow{\sim} F$. In general, the ring $W(A)$ is equal to the inverse limit of the rings $W_n(A)$ of Witt vectors of length n. But rather than forming the limit, we shall consider the limit system of rings $W.(A)$ as a pro-ring. There is a natural map of pro-rings $F\colon W.(A) \to W._{-1}(A)$, called the Frobenius, and a natural map of $W.(A)$-modules $V\colon F_* W._{-1}(A) \to W.(A)$, called the Verschiebung. The former is given, as the ring structure, by certain polynomials in the coordinates; the latter is given by $V(a_0, \dots, a_{n-2}) = (0, a_0, \dots, a_{n-2})$, and $FV = p$. Finally, we note that if (A, M) is a log ring, then the composite

$$M \xrightarrow{\alpha} A \xrightarrow{[\ ]} W.(A)$$

makes $(W.(A), M)$ a pro-log ring.

There is a natural way to combine differential forms and Witt vectors; the result is called the de Rham-Witt complex. It was considered first for $\mathbb{F}_p$-algebras by Bloch-Deligne-Illusie [3, 23] in connection with the crystalline cohomology of Berthelot-Grothendieck [2]. A generalization to log-$\mathbb{F}_p$-algebras was constructed by Hyodo-Kato [22]. The following extension to log-$\mathbb{Z}_{(p)}$-algebras was obtained in collaboration with Ib Madsen [19, 20]: Let (A, M) be a log ring such that A is a $\mathbb{Z}_{(p)}$-algebra with p odd. A *Witt complex* over (A, M) is:

(i) a pro-log differential graded ring $(E^*_\cdot, M_E)$ and a map of pro-log rings

$$\lambda\colon (W_\cdot(A), M) \to (E^0_\cdot, M_E);$$

(ii) a map of pro-log graded rings

$$F\colon E^*_\cdot \to E^*_{\cdot-1},$$

such that $\lambda F = F\lambda$ and such that

$$\begin{aligned} Fd\log_n a &= d\log_{n-1} a, && \text{for all } a \in M, \\ Fd\,\lambda[a]_n &= \lambda[a]_{n-1}^{p-1} d\,\lambda[a]_{n-1}, && \text{for all } a \in A; \end{aligned}$$

(iii) a map of pro-graded modules over the pro-graded ring $E^*_\cdot$,

$$V\colon F_*\, E^*_{\cdot-1} \to E^*_\cdot,$$

such that $\lambda V = V\lambda$, $FV = p$ and $FdV = d$.

A map of Witt complexes over (A, M) is a map of pro-log differential graded rings which commutes with the maps λ, F and V. Standard category theory shows that there exists a universal Witt complex over (A, M). This, by defintion, is the de Rham-Witt complex $W_\cdot\Omega^*_{(A,M)}$. (The canonical maps $W_\cdot(A) \to W_\cdot\Omega^0_{(A,M)}$ and $\Omega^*_{(A,M)} \to W_1\,\Omega^*_{(A,M)}$ are isomorphisms, so the construction really does combine differential forms and Witt vectors.) We lift the logarithmic derivative to a map

$$K^M_q(K) \to W_n\,\Omega^q_{(V,M)}$$

which to the symbol $\{a_1, \dots, a_q\}$ associates $d\log_n a_1 \dots d\log_n a_q$. This trace map better captures the Milnor K-groups. Indeed, the following result was obtained in collaboration with Thomas Geisser [14]:

Theorem 2.1 *Suppose that $\mu_{p^v} \subset K$ and that k is separably closed. Then the trace map induces an isomorphism of pro-abelian groups*

$$K^M_q(K)/p^v \xrightarrow{\sim} \left(W_\cdot\Omega^q_{(V,M)}/p^v\right)^{F=1}.$$

To prove this, we first show that $W_n\,\Omega^q_{(V,M)}/p$ has a (non-canonical) k-vector space structure and find an explicit basis. The dimension is

$$\dim_k\left(W_n\,\Omega^q_{(V,M)}/p\right) = \binom{r+1}{q}\, e \sum_{s=0}^{n-1} p^{rs},$$

where $|k : k^p| = p^r$ and e the ramification index of K. It is not difficult to see that this is an upper bound for the dimension. The proof that it is also a lower bound is more involved and uses a formula for the de Rham-Witt complex of a polynomial extension by Madsen and the author [19]. We then evaluate the kernel of $1 - F$ and compare with the calculation of $K^M_q(K)/p$ by Kato [24, 4]. The assumption

that the residue field k be separably closed is not essential. In the general case, one instead has a short-exact sequence

$$0 \to \left(W.\,\Omega^{q-1}_{(V,M)} \otimes \mu_{p^v}\right)_{F=1} \to K^M_q(K)/p^v \to \left(W.\,\Omega^q_{(V,M)}/p^v\right)^{F=1} \to 0,$$

where the superscript (resp. subscript) "$F = 1$" indicates Frobenius invariants (resp. coinvariants).

We discuss a global version of theorem 2.1. Let V_0 be a henselian discrete valuation ring with quotient field K_0 of characteristic zero and *perfect* residue field k_0 of odd characteristic p. Let $\mathfrak{X}$ be a smooth V_0-scheme, and let i (resp. j) denote the inclusion of the special (resp. generic) fiber as in the cartesian diagram

$$\begin{array}{ccccc} X & \overset{j}{\hookrightarrow} & \mathfrak{X} & \overset{i}{\hookleftarrow} & Y \\ \downarrow & & \downarrow{\scriptstyle f} & & \downarrow \\ \operatorname{Spec} K_0 & \hookrightarrow & \operatorname{Spec} V_0 & \hookleftarrow & \operatorname{Spec} k_0. \end{array}$$

Suppose that $\mu_{p^v} \subset K_0$. Then the proof of theorem 2.1 shows that there is a short-exact sequence of sheaves of pro-abelian groups on Y for the étale topology

$$0 \to i^* R^q j_* \mathbb{Z}/p^v(q) \to i^*(W.\,\Omega^q_{(\mathfrak{X},M)}/p^v) \xrightarrow{1-F} i^*(W.\,\Omega^q_{(\mathfrak{X},M)}/p^v) \to 0.$$

The left hand term is the sheaf of p-adic vanishing cycles.

3. The cyclotomic trace

We now turn to Quillen K-theory. The analog of the logarithmic derivative is the topological Dennis trace with values in topological Hochschild homology,

$$K_*(\mathcal{C}) \to \mathrm{THH}_*(\mathcal{C}),$$

defined by Bökstedt [5]. It is a refinement of earlier trace maps by Dennis [9] and Waldhausen [35]. We will use a variant of the construction due to Dundas-McCarthy [10, 27] that can be applied to a category with cofibrations and weak equivalences in the sense of Waldhausen [36]. The category $\mathcal{C}$ we consider is the category of bounded chain complexes of finitely generated projective V-modules. The cofibrations are the degree-wise monomorphisms, and the weak equivalences are the chain maps $C \to C'$ such that $K \otimes_V C \to K \otimes_V C'$ is a quasi-isomorphism. The K-theory of this category is canonically isomorphic to Quillen's K-theory of the field K. We showed in [20] that the groups

$$\mathrm{THH}_*(V|K) = \mathrm{THH}_*(\mathcal{C})$$

form a log differential graded ring with underlying log ring (V, M), where the structure map $d\log$ is given by the composite

$$M = V \cap K^* \xrightarrow{\ell} K_1(K) \to \mathrm{THH}_1(V|K).$$

The canonical map from the de Rham complex

$$\Omega^q_{(V,M)} \to \mathrm{THH}_q(V|K)$$

is compatible with the trace maps and is an isomorphism, if $q \leq 2$. The topological Dennis trace, again, is far from injective. This can be rectified by a construction which, in retrospect, can be seen as incorporating Witt vectors. The result is the cyclotomic trace of Bökstedt-Hsiang-Madsen [6] which we now recall. The reader is referred to [20, 15, 10] for details.

The topological Dennis trace, we recall, is defined as the map of homotopy groups induced from a continuous map of spaces

$$K(\mathcal{C}) \to \mathrm{THH}(\mathcal{C}).$$

As a consequence of Connes' theory of cyclic sets, the right hand space is equipped with a continuous action by the circle group $\mathbb{T}$. Moreover, the image of the trace map is point-wise fixed by the $\mathbb{T}$-action. Let

$$\mathrm{TR}^n(\mathcal{C}; p) = \mathrm{THH}(\mathcal{C})^{C_{p^{n-1}}}$$

be the subset fixed by the subgroup $C_{p^{n-1}} \subset \mathbb{T}$ of the indicated order. It turns out that, as n and q varies, the homotopy groups

$$\mathrm{TR}^n_q(V|K; p) = \pi_q(\mathrm{TR}^n(\mathcal{C}; p))$$

form a Witt complex over (V, M); see [21, 18, 20]. The map F is induced from the obvious inclusion map, and the map V is the accompanying transfer map. The structure maps in the limit system and the map λ, however, are more difficult to define. The former was defined in [6] and the latter in [21]. The topological Dennis trace induces a map of pro-abelian groups

$$K_q(K) \to \mathrm{TR}^{\cdot}_q(V|K; p).$$

This is the cyclotomic trace. It takes values in the subset fixed by the Frobenius operator. The canonical map

$$W_n\Omega^q_{(V,M)} \to \mathrm{TR}^n_q(V|K; p)$$

is compatible with the trace maps from Milnor K-theory and Quillen K-theory, respectively, and is an isomorphism, if $q \leq 2$. The following is a combination of results obtained in collaboration with Thomas Geisser [15] and Ib Madsen [21, 20].

Theorem 3.1 *Suppose that k is separably closed. Then the cyclotomic trace induces an isomorphism of pro-abelian groups*

$$K_q(K, \mathbb{Z}/p^v) \xrightarrow{\sim} \mathrm{TR}^{\cdot}_q(V|K; p, \mathbb{Z}/p^v)^{F=1}.$$

We briefly outline the steps in the proof: We proved in [15] that the sequence

$$0 \to K_q(k, \mathbb{Z}/p^v) \to \mathrm{TR}_q^\cdot(k; p, \mathbb{Z}/p^v) \xrightarrow{1-F} \mathrm{TR}_q^\cdot(k; p, \mathbb{Z}/p^v) \to 0$$

is exact. This uses [4, 16, 18]. Given this, the theorem by McCarthy [28] that for nilpotent extensions, relative K-theory and relative topological cyclic homology agree, and the continuity results of Suslin [32] for K-theory and Madsen and the author [21] for TR show that also the sequence

$$0 \to K_q(V, \mathbb{Z}/p^v) \to \mathrm{TR}_q^\cdot(V; p, \mathbb{Z}/p^v) \xrightarrow{1-F} \mathrm{TR}_q^\cdot(V; p, \mathbb{Z}/p^v) \to 0$$

is exact. Theorem 3.1 follows by comparing the localization sequence of Quillen [30]

$$\cdots \to K_q(k, \mathbb{Z}/p^v) \xrightarrow{i^!} K_q(V, \mathbb{Z}/p^v) \xrightarrow{j_*} K_q(K, \mathbb{Z}/p^v) \to \cdots$$

to the corresponding sequence by Madsen and the author [20]

$$\cdots \to \mathrm{TR}_q^n(k; p, \mathbb{Z}/p^v) \xrightarrow{i^!} \mathrm{TR}_q^n(V; p, \mathbb{Z}/p^v) \xrightarrow{j_*} \mathrm{TR}_q(V|K; p, \mathbb{Z}/p^v) \to \cdots .$$

Again, the assumption in the statement of theorem 3.1 that the residue field k be separably closed is not essential. The general statement will be given below. It is also not necessary for theorem 3.1 to assume that V be of geometric type.

4. The Tate spectral sequence

If G is a finite group and X a G-space, it is usually not possible to evaluate the groups $\pi_*(X^G)$ from knowledge of the G-modules $\pi_*(X)$. At first glance, this is the problem that one faces in evaluating the groups

$$\mathrm{TR}_q^n(\mathcal{C}; p) = \pi_q\big(\mathrm{THH}(\mathcal{C})^{C_{p^{n-1}}}\big).$$

However, the mapping fiber of the structure map $\mathrm{TR}^n(\mathcal{C}; p) \to \mathrm{TR}^{n-1}(\mathcal{C}; p)$, it turns out, is given by the Borel construction $\mathbb{H}.(C_{p^{n-1}}, \mathrm{THH}(\mathcal{C}))$ whose homotopy groups are the abutment of a (first quadrant) spectral sequence

$$E^2_{s,t} = H_s(C_{p^{n-1}}, \mathrm{THH}_t(\mathcal{C})) \Rightarrow \pi_{s+t}\mathbb{H}.(C_{p^{n-1}}, \mathrm{THH}(\mathcal{C})).$$

This suggests that the groups $\mathrm{TR}_q^n(\mathcal{C}; p)$ can be evaluated inductively starting from the case $n = 1$. However, it is generally difficult to carry out the induction step. In addition, the absence of a multiplicative structure makes the spectral sequence above difficult to solve. The main vehicle to overcome these problems, first employed by Bökstedt-Madsen in [7], is the following diagram of fiber sequences

$$\begin{array}{ccccc}
\mathbb{H}.(C_{p^{n-1}}, \mathrm{THH}(\mathcal{C})) & \longrightarrow & \mathrm{TR}^n(\mathcal{C}; p) & \longrightarrow & \mathrm{TR}^{n-1}(\mathcal{C}; p) \\
\| & & \downarrow{\scriptstyle \Gamma} & & \downarrow{\scriptstyle \hat{\Gamma}} \\
\mathbb{H}.(C_{p^{n-1}}, \mathrm{THH}(\mathcal{C})) & \longrightarrow & \mathbb{H}^\cdot(C_{p^{n-1}}, \mathrm{THH}(\mathcal{C})) & \longrightarrow & \hat{\mathbb{H}}(C_{p^{n-1}}, \mathrm{THH}(\mathcal{C}))
\end{array}$$

together with a multiplicative (upper half-plane) spectral sequence

$$E^2_{s,t} = \hat{H}^{-s}(C_{p^{n-1}}, \mathrm{THH}_t(\mathcal{C})) \Rightarrow \pi_{s+t}\hat{\mathbb{H}}(C_{p^{n-1}}, \mathrm{THH}(\mathcal{C}))$$

starting from the Tate cohomology of the (trivial) $C_{p^{n-1}}$-module $\mathrm{THH}_t(\mathcal{C})$. The lower fiber sequence is the Tate sequence; see Greenlees and May [17] or [20]. In favorable cases, the maps Γ and $\hat{\Gamma}$ induce isomorphisms of homotopy groups in non-negative degrees. Indeed, this is true in the case at hand (if k is perfect). The differential structure of the spectral sequence

$$E^2_{s,t} = \hat{H}^{-s}(C_{p^{n-1}}, \mathrm{THH}_t(V|K, \mathbb{Z}/p)) \Rightarrow \pi_{s+t}(\hat{\mathbb{H}}(C_{p^{n-1}}, \mathrm{THH}(V|K)), \mathbb{Z}/p)$$

was determined in collaboration with Ib Madsen [20] in the case where the residue field k is perfect. This is the main calculational result of the work reported here. The following result, for perfect k, is a rather immediate consequence. The extension to non-perfect k is given in [19].

Theorem 4.1 *Suppose that $\mu_{p^v} \subset K$. Then the canonical map is an isomorphism of pro-abelian groups*

$$W.\,\Omega^*_{(V,M)} \otimes_{\mathbb{Z}} S_{\mathbb{Z}/p^v}(\mu_{p^v}) \xrightarrow{\sim} \mathrm{TR}^{\cdot}_*(V|K; p, \mathbb{Z}/p^v).$$

We can now state the general version of theorem 3.1 which does not require that the residue field k be separably closed. The second tensor factor on the left hand side in the statement of theorem 4.1 is the symmetric algebra on the $\mathbb{Z}/p^v$-module μ_{p^v}, which is free of rank one. Spelling out the statement for the group in degree q, we get an isomorphism of pro-abelian groups

$$\bigoplus_{s\geq 0} W.\,\Omega^{q-2s}_{(V,M)} \otimes \mu_{p^v}^{\otimes s} \xrightarrow{\sim} \mathrm{TR}^{\cdot}_q(V|K; p, \mathbb{Z}/p^v).$$

In the case of a separably closed residue field, theorem 3.1 idenfies the Frobenius fixed set of the common pro-abelian group with $K_q(K, \mathbb{Z}/p^v)$. In the general case, one has instead a short-exact sequence

$$0 \to \bigoplus_{s\geq 1} \left(W.\,\Omega^{q+1-2s}_{(V,M)} \otimes \mu_{p^v}^{\otimes s}\right)_{F=1} \to K_q(K, \mathbb{Z}/p^v) \to \bigoplus_{s\geq 0} \left(W.\,\Omega^{q-2s}_{(V,M)} \otimes \mu_{p^v}^{\otimes s}\right)^{F=1} \to 0,$$

valid for all integers q. (There is a similar sequence for the topological cyclic homology group $\mathrm{TC}^{\cdot}_q(V|K; p, \mathbb{Z}/p^v)$ [20] which includes the summand "$s = 0$" on the left.) Comparing with the general version of theorem 2.1, we obtain the following result promised earlier [20, 14].

Theorem 4.2 *Suppose that $\mu_{p^v} \subset K$. Then the canonical map*

$$K^M_*(K) \otimes_{\mathbb{Z}} S_{\mathbb{Z}/p^v}(\mu_{p^v}) \xrightarrow{\sim} K_*(K, \mathbb{Z}/p^v)$$

is an isomorphism.

5. Galois descent

We now assume that the residue field k be perfect. In homotopy theoretic terms, theorem 4.1 states that the pro-spectrum $\mathrm{TR}^{\cdot}(V|K;p)$ is equivalent to the (-1)-connected cover of its localization with respect to complex periodic K-theory, see [8]. This suggests the possibility of completely understanding the homotopy type of this pro-spectrum. We expect that this, in turn, is closely related to the following question. Let $\bar{K}$ be an algebraic closure of K with Galois group G_K, and let $\bar{V}$ be the integral closure of V in $\bar{K}$. (The ring $\bar{V}$ is a valuation ring with value group the additive group of rational numbers.)

Conjecture 5.1 *If k is perfect then for all $q > 0$, the canonical map*

$$\mathrm{TR}^{\cdot}_q(V|K;p,\mathbb{Q}_p/\mathbb{Z}_p) \to \mathrm{TR}^{\cdot}_q(\bar{V}|\bar{K};p,\mathbb{Q}_p/\mathbb{Z}_p)^{G_K}$$

be an isomorphism of pro-abelian groups and that the higher continuous cohomology groups $H^i_{\mathrm{cont}}(G_K,\mathrm{TR}^n_q(\bar{V}|\bar{K};p,\mathbb{Q}_p/\mathbb{Z}_p))$ vanish.

It follows from Tate [33] that the groups $H^i_{\mathrm{cont}}(G_K,\mathrm{TR}^n_q(\bar{V}|\bar{K};p,\mathbb{Q}_p))$ vanish for $i \geq 0$ and $q > 0$. One may hope that these methods will help shed some light on the structure of the groups $H^i_{\mathrm{cont}}(G_K,\mathrm{TR}^n_q(\bar{V}|\bar{K};p,\mathbb{Q}_p/\mathbb{Z}_p))$. We now describe the structure of these G_K-modules; proofs will appear elsewhere.

The group $\mathrm{TR}^n_q(\bar{V}|\bar{K};p,\mathbb{Q}_p/\mathbb{Z}_p)$ is divisible, if $q > 0$, and uniquely divisible, if $q > 0$ and even. The Tate module $T_p\,\mathrm{TR}^n_1(\bar{V}|\bar{K};p)$ is a free module of rank one over $\mathrm{TR}^n_0(\bar{V}|\bar{K};p,\mathbb{Z}_p)$, and the canonical map an isomorphism:

$$S_{\mathrm{TR}^n_0(\bar{V}|\bar{K};p,\mathbb{Z}_p)}(T_p\,\mathrm{TR}^n_1(\bar{V}|\bar{K};p)) \xrightarrow{\sim} \mathrm{TR}^n_*(\bar{V}|\bar{K};p,\mathbb{Z}_p)$$

(note that $\mathrm{TR}^n_q(\bar{V}|\bar{K};p,\mathbb{Q}_p/\mathbb{Z}_p) \xrightarrow{\sim} \mathrm{TR}^n_q(\bar{V}|\bar{K};p,\mathbb{Z}_p)\otimes\mathbb{Q}_p/\mathbb{Z}_p$). We note the formal analogy with the results on $K_*(\bar{K})$ by Suslin [31, 32].

The structure of the ring $\mathrm{TR}^n_0(\bar{V}|\bar{K};p,\mathbb{Z}_p) = W_n(\bar{V})^{\wedge}$ is well-understood (unlike that of $W_n(V)$): Following Fontaine [13], we let $R_{\bar{V}}$ be the inverse limit of the diagram $\bar{V}/p \leftarrow \bar{V}/p \leftarrow \cdots$ with the Frobenius as structure map. This is a perfect $\mathbb{F}_p$-algebra and an integrally closed domain whose quotient field is algebraically closed. There is a surjective ring homomorphism $\theta_n\colon W(R_{\bar{V}}) \twoheadrightarrow W_n(\bar{V})^{\wedge}$ whose kernel is a principal ideal. If $\epsilon = \{\epsilon^{(v)}\}_{v\geq 1}$ is a compatible sequence of primitive p^{v-1}st roots of unity considered as an element of $R_{\bar{V}}$, and if ϵ_n is the unique p^nth root of ϵ, then $([\epsilon]-1)/([\epsilon_n]-1)$ is a generator. Moreover, as n varies, the maps θ_n constitute a map of pro-rings compatible with the Frobenius maps.

The Bott element $b_{\epsilon,n} \in T_p\,\mathrm{TR}^n_1(\bar{V}|\bar{K};p)$ determined by the sequence ϵ is not a generator (so the statement of theorem 4.1 is not valid for $\bar{K}$). Instead there is a generator $\alpha_{\epsilon,n}$ such that $b_{\epsilon,n} = ([\epsilon_n]-1)\cdot\alpha_{\epsilon,n}$. The structure map of the pro-abelian group $T_p\,\mathrm{TR}^{\cdot}_1(\bar{V}|\bar{K};p)$ (resp. the Frobenius) takes $\alpha_{\epsilon,n}$ to $([\epsilon_{n-1}]-1)/([\epsilon_n]-1)\cdot\alpha_{\epsilon,n-1}$ (resp. to $\alpha_{\epsilon,n-1}$), and the action of the Galois group is given by

$$\alpha^{\sigma}_{\epsilon,n} = \chi(\sigma)\frac{[\epsilon_n]-1}{[\epsilon^{\sigma}_n]-1}\cdot\alpha_{\epsilon,n},$$

where $\chi\colon G_K \to \mathrm{Aut}(\mu_{p^\infty}) = \mathbb{Z}_p^*$ is the cyclotomic character.

Acknowledgments The research reported here was supported in part by grants from the National Science Foundation and by an Alfred P. Sloan Fellowship.

References

[1] A. A. Beilinson, *Height pairing between algebraic cycles*, K-theory, arithmetic and geometry (Moscow, 1984–1986), Lecture Notes in Math., vol. 1289, Springer-Verlag, 1987, 1–25.

[2] P. Berthelot, *Cohomologie cristalline des schemas de caracteristique* $p > 0$, Lecture Notes in Math., vol. 407, Springer-Verlag, 1974.

[3] S. Bloch, *Algebraic K-theory and crystalline cohomology*, Publ. Math. I.H.E.S. **47** (1977), 187–268.

[4] S. Bloch and K. Kato, *p-adic etale cohomology*, Publ. Math. IHES **63** (1986), 107–152.

[5] M. Bökstedt, *Topological Hochschild homology*, Preprint 1985, Universität Bielefeld.

[6] M. Bökstedt, W.-C. Hsiang, and I. Madsen, *The cyclotomic trace and algebraic K-theory of spaces*, Invent. Math. **111** (1993), 465–540.

[7] M. Bökstedt and I. Madsen, *Topological cyclic homology of the integers*, K-theory (Strasbourg, 1992), Astérisque, vol. 226, 1994, 57–143.

[8] A. K. Bousfield, *The localization of spectra with respect to homology*, Topology **18** (1979), 257–281.

[9] K. Dennis, *Algebraic K-theory and Hochschild homology*, Algebraic K-theory, Evanston, IL., 1976 (unpublished lecture).

[10] B. I. Dundas and R. McCarthy, *Topological Hochschild homology of ring functors and exact categories*, J. Pure Appl. Alg. **109** (1996), 231–294.

[11] J. L. Dupont, *Algebra of polytopes and homology of flag complexes*, Osaka J. Math. **19** (1982), 599–641.

[12] J. L. Dupont and C.-H. Sah, *Scissors congruences, II*, J. Pure Appl. Alg. **25** (1982), 159–195.

[13] J.-M. Fontaine, *Le corps des périodes p-adiques*, Périodes p-adiques (Séminaire de Bures, 1988), Astérisque, vol. 223, 1994, 59–111.

[14] T. Geisser and L. Hesselholt, *On the K-theory of a henselian discrete valuation field with non-perfect residue field*, Preprint 2002.

[15] ———, *Topological cyclic homology of schemes*, K-theory (Seattle, 1997), Proc. Symp. Pure Math., vol. 67, 1999, 41–87.

[16] T. Geisser and M. Levine, *The K-theory of fields in characteristic p*, Invent. Math. **139** (2000), 459–493.

[17] J. P. C. Greenlees and J. P. May, *Generalized Tate cohomology*, vol. 113, Mem. Amer. Math. Soc., no. 543, 1995.

[18] L. Hesselholt, *On the p-typical curves in Quillen's K-theory*, Acta Math. **177** (1997), 1–53.

[19] L. Hesselholt and I. Madsen, *On the de Rham-Witt complex in mixed characteristic*, Preprint 2002.

[20] ———, *On the K-theory of local fields*, Ann. Math. (to appear).
[21] ———, *On the K-theory of finite algebras over Witt vectors of perfect fields*, Topology **36** (1997), 29–102.
[22] O. Hyodo and K. Kato, *Semi-stable reduction and crystalline cohomology with logarithmic poles*, Périodes p-adiques, Asterisque, vol. 223, 1994, 221–268.
[23] L. Illusie, *Complexe de de Rham-Witt et cohomologie cristalline*, Ann. Scient. Éc. Norm. Sup. (4) **12** (1979), 501–661.
[24] K. Kato, *Galois cohomology of complete discrete valuation fields*, Algebraic K-theory, Part II (Oberwolfach, 1980), Lecture Notes in Math., vol. 967, Springer, Berlin-New York, 1982, 215–238.
[25] ———, *Logarithmic structures of Fontaine-Illusie*, Algebraic analysis, geometry, and number theory, Proceedings of the JAMI Inaugural Conference (Baltimore, 1988), Johns Hopkins Univ. Press, Baltimore, MD, 1989, 191–224.
[26] S. Licthenbaum, *Values of zeta-functions at non-negative integers*, Number theory, Lecture Notes in Math., vol. 1068, Springer-Verlag, 1983, 127–138.
[27] R. McCarthy, *The cyclic homology of an exact category*, J. Pure Appl. Alg. **93** (1994), 251–296.
[28] ———, *Relative algebraic K-theory and topological cyclic homology*, Acta Math. **179** (1997), 197–222.
[29] J. Milnor, *Algebraic K-theory and quadratic forms*, Invent. Math. **9** (1970), 318–344.
[30] D. Quillen, *Higher algebraic K-theory I*, Algebraic K-theory I: Higher K-theories (Battelle Memorial Inst., Seattle, Washington, 1972), Lecture Notes in Math., vol. 341, Springer-Verlag, 1973.
[31] A. A. Suslin, *On the K-theory of algebraically closed fields*, Invent. Math. **73** (1983), 241–245.
[32] ———, *On the K-theory of local fields*, J. Pure Appl. Alg. **34** (1984), 304–318.
[33] J. Tate, *p-divisible groups*, Proc. conf. local fields (Driebergen, 1966), Springer-Verlag, 1967, 158–183.
[34] V. Voevodsky, *The Milnor conjecture*, Preprint 1996, Max Planck Institut, Bonn.
[35] F. Waldhausen, *Algebraic K-theory of topological spaces. II*, Algebraic topology (Aarhus, 1978), Lecture Notes in Math., vol. 763, Springer-Verlag, 1979, 356–394.
[36] ———, *Algebraic K-theory of spaces*, Algebraic and geometric topology (New Brunswick, N.J., 1983), Lecture Notes in Math., vol. 1126, Springer-Verlag, 1985, 318–419.

ICM 2002 · Vol. II · 427–436

Symplectic Sums and Gromov-Witten Invariants

Eleny-Nicoleta Ionel*

Abstract

Gromov-Witten invariants of a symplectic manifold are a count of holomorphic curves. We describe a formula expressing the GW invariants of a symplectic sum $X\#Y$ in terms of the relative GW invariants of X and Y. This formula has several applications to enumerative geometry. As one application, we obtain new relations in the cohomology ring of the moduli space of complex structures on a genus g Riemann surface with n marked points.

2000 Mathematics Subject Classification: 57R17, 53D45, 14N35.

1. Gromov-Witten invariants

A symplectic structure on a closed smooth manifold X^{2N} consists of a closed, non-degenerate 2-form ω. Gromov's idea [8] was that one could obtain information about the symplectic structure on X by studying holomorphic curves. For that one needs to introduce an almost complex structure, which is an endomorphism $J \in End(TX)$ with $J^2 = -Id$. Such a J is compatible with ω if the bilinear form $g(v,w) = \omega(v,Jw)$ defines a Riemannian metric on TX. For a fixed symplectic structure, the space of compatible almost complex structures is a nonempty, contractible space.

One then considers the moduli space of J-holomorphic maps from Riemann surfaces into X. Constraints are imposed on the maps, requiring the domain to have a certain form and the image to pass through geometric representatives of fixed homology classes in X. When the right number of constraints are chosen there will be finitely many maps satisfying those constraints; the (oriented) count of these maps will give the corresponding Gromov-Witten invariant. In general, there are several technical difficulties one must overcome to get a well-defined Gromov-Witten invariant. The foundations of this theory began with [8], [24], [25] and have been developed since then by the efforts of a large group of mathematicians (see,

*Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA. E-mail: ionel@math.wisc.edu

for example, the references in [15] and [22]). Here we present a brief overview of the technical setup.

Consider (X, ω) a symplectic manifold. For each compatible almost complex structure J and perturbation ν one considers maps $f : C \to X$ from a genus g Riemann surface C with n marked points which satisfy the pseudo-holomorphic map equation $\overline{\partial} f = \nu$ and represent a fixed homology class $A = [f] \in H_2(X)$. The set of such maps (modulo reparametrizations), together with their limits, forms the compact space of stable maps $\overline{\mathcal{M}}_{g,n}(X, A)$. For each stable map $f : C \to X$, the domain determines a point in the Deligne-Mumford moduli space $\overline{\mathcal{M}}_{g,n}$ of genus g Riemann surfaces with n marked points (see also §3). The evaluation at each marked point determines a point in X. All together, this gives a natural map

$$\overline{\mathcal{M}}_{g,n}(X, A) \longrightarrow \overline{\mathcal{M}}_{g,n} \times X^n.$$

For generic (J, ν) the image of this map carries a fundamental homology class $[GW_{X,A,g,n}]$ which is defined to be the Gromov-Witten invariant of (X, ω). The dimension of this homology class, given by an index computation, is

$$\dim \overline{\mathcal{M}}_{g,n}(X, A) = 2c_1(TX)A + (\dim X - 6)(1 - g) + 2n.$$

A cobordism argument shows that the homology class $[GW_{X,A,g,n}]$ is independent of generic (J, ν) and moreover depends only on the isotopy class of the symplectic form ω. Frequently, the Gromov-Witten invariant is thought of as a collection of numbers obtained by evaluating the homology class $[GW_{X,A,g,n}]$ on a basis of the dual cohomology group. For complex algebraic manifolds these symplectic invariants can also be defined by algebraic geometry, and in important cases the invariants are the same as the counts of curves that are the subject of classical enumerative algebraic geometry.

The next important question is to find effective ways of computing the GW invariants. One useful technique is the method of 'splitting the domain'. Anytime we have a relation in the cohomology of $\overline{\mathcal{M}}_{g,n}$ it pulls back to a relation (sometimes trivial) between the GW invariants of a symplectic manifold X. As an example, suppose that the constraints imposed on the domain of the holomorphic curves are boundary classes in $H^*(\overline{\mathcal{M}}_{g,n})$ (as defined in section 3 below). One then obtains recursive relations which relate such GW invariant to invariants of lower degree or genus. This method was first used by Kontsevich and Ruan-Tian [25] to determine recursively the genus 0 invariants of the projective spaces $\mathbb{P}^n$. These recursive relations follow from the observation that in the Deligne-Mumford space $\overline{\mathcal{M}}_{0,4} \cong \mathbb{P}^1$ each boundary class corresponds to a point, and are thus all homologous to each other.

In joint work with Thomas H. Parker, the author established a general formula describing the behavior of GW invariants under the operation of 'splitting the target' ([14], [15], [16]). Because we work in the context of symplectic manifolds the natural splitting of the target is the one associated with the symplectic cut operation and its inverse, the symplectic sum. The next section describes the symplectic sum operation and the main ingredients entering the sum formula for GW invariants.

2. Symplectic sums

The operation of symplectic sum is defined by gluing along codimension two submanifolds (see [7], [21]). Specifically, let X be a symplectic manifold with a codimension two symplectic submanifold V. Given a similar pair (Y, V) with a symplectic identification between the two copies of V and a complex anti-linear isomorphism between the normal bundles $N_X V$ and $N_Y V$ of V in X and in Y we can form the symplectic sum $X \#_V Y$.

Perhaps it is in more natural to describe the symplectic sum not as a single manifold but as a family $Z \to D$ over the disk depending on a parameter $\lambda \in D$. For $\lambda \neq 0$ the fibers Z_λ are smooth and symplectically isotopic to $X \#_V Y$ while the central fiber Z_0 is the singular manifold $X \cup_V Y$. In a neighborhood of V the total space Z is $N_X V \oplus N_Y V$ and the fiber Z_λ is defined by the equation $xy = \lambda$ where x and y are coordinates in the normal bundles $N_X V$ and $N_Y V \cong (N_X V)^*$. The fibration $Z \to D$ extends away from V as the disjoint union of $X \times D$ and $Y \times D$.

Our overall strategy for proving the symplectic sum formula for GW invariants [16] is to relate the pseudo-holomorphic maps into Z_λ for λ small to pseudo-holomorphic maps into Z_0. One expects the stable maps into the sum to be pairs of stable maps into the two sides which match in the middle. A sum formula thus requires a count of stable maps in X that keeps track of how the curves intersect V.

So the first step is to construct Gromov-Witten invariants for a symplectic manifold (X, ω) relative to a codimension two symplectic submanifold V. These invariants were introduced in a separate paper with Thomas H. Parker [15] and were designed for use in symplectic sum formulas. Of course, before speaking of stable maps one must extend the almost complex structure J and the perturbation ν to the symplectic sum. To ensure that there is such an extension we require that the pair (J, ν) be V-compatible. The precise definition is given in section §6 of [15], but in particular for such pairs V is a J-holomorphic submanifold — something which is not true for generic J. The relative invariant gives counts of stable maps for these special V-compatible pairs. Such counts are in general different from those associated with the absolute GW invariants described in the first section of this note.

Restricting to V-compatible pairs has repercussions. Any pseudo-holomorphic map $f : C \to V$ into V then automatically satisfies the pseudo-holomorphic map equation into X. So for V-compatible (J, ν), stable maps may have domain components whose image lies entirely in V, so they are far from being transverse to V. Worse, the moduli spaces of such maps can have dimension *larger* than the dimension of $\mathcal{M}_{g,n}(X, A)$. We circumvent these difficulties by restricting attention to the stable maps which *have no components mapped entirely into* V. Such 'V-regular' maps intersect V in a finite set of points with multiplicity. After numbering these points, the space of V-regular maps separates into components labeled by vectors $s = (s_1, \ldots, s_\ell)$, where ℓ is the number of intersection points and s_k is the multiplicity of the k^{th} intersection point. Each (irreducible) component $\mathcal{M}^V_{g,n,s}(X, A)$ of V-regular stable maps is an orbifold; its dimension depends of g, n, A and on the vector of multiplicities s.

Next key step is to show that the space of V-regular maps carries a fundamental homology class. For this we construct an orbifold compactification $\overline{\mathcal{M}}^V_{g,n,s}(X,A)$, the space of V-stable maps. The relative invariants are then defined in exactly the same way as the GW invariants. We consider the natural map

$$\overline{\mathcal{M}}^V_{g,n,s}(X,A) \to \overline{\mathcal{M}}_{g,n+\ell} \times X^n \times V^\ell. \tag{2.1}$$

The new feature is the last factor (the evaluation at the ℓ points of contact with V) which allows us to constrain how the images of the maps intersect V. Thus the relative invariants give counts of V-stable maps with constraints on the complex structure of the domain, the images of the marked points, and the geometry of the intersection with V. There is one more complication: to be useful for a symplectic sum formula, the relative invariant should record the homology class of the curve in $X \setminus V$ rather than in X. This requires keeping track of some additional homology data which is intertwined with the intersection data, as explained in [15].

We now return to the discussion of the symplectic sum formula. As previously mentioned, the overall strategy is to relate the pseudo-holomorphic maps into Z_0, which are simply maps into X and Y which match along V, with pseudo-holomorphic maps into Z_λ for λ close to zero. For that we consider sequences of stable maps into the family Z_λ of symplectic sums as the 'neck size' $\lambda \to 0$. These limit to maps into the singular manifold $Z_0 = X \cup_V Y$. A more careful look reveals several features of the limit maps.

First of all, if the limit map $f_0 : C_0 \to Z_0$ has no components in V then f_0 has matching intersection with V on X and Y side. For such a limit map f_0 all its intersection points with V are nodes of the domain C_0. Ordering this nodes we obtain a sequence of multiplicities $s = (s_1, \ldots, s_\ell)$ along V. But it turns out that the squeezing process is not injective in general. For a fixed $\lambda \neq 0$ there are $|s| = s_1 \cdot \ldots \cdot s_\ell$ many stable maps into Z_λ close to f_0.

Second, connected curves in Z_λ can limit to curves whose restrictions to X and Y are not connected. For that reason the GW invariant, which counts stable curves from a connected domain, is not the appropriate invariant for expressing a sum formula. Instead one should work with the 'Gromov-Taubes' invariant GT, which counts stable maps from domains that need not be connected. Thus we seek a formula of the general form

$$GT_{X\#_V Y} \;=\; GT^V_X \,*\, GT^V_Y \tag{2.2}$$

where $*$ is the operation that adds up the ways curves on the X and Y sides match and are identified with curves in Z_λ. That necessarily involves keeping track of the multiplicities s and the homology classes. It also involves accounting for the limit maps which have components in V; such maps are not counted by the relative invariant and hence do not contribute to the left side of (2.2).

Finally, we need to consider limit maps which have components mapped entirely in V. We deal with that possibility by squeezing the neck not in one region, but several regions. As a result, the formula (2.2) in general has an extra term

called the S-matrix which keeps track of how the genus, homology class, and intersection points with V change as the images of stable maps pass through the neck region. One sees these quantities changing abruptly as the map passes through the neck — the maps are "scattered" by the neck. The scattering occurs when some of the stable maps contributing to the GT invariant of Z_λ have components that lie entirely in V in the limit as $\lambda \to 0$. Those maps are not V-regular, so are not counted in the relative invariants of X or Y. But this complication can be analyzed and related to the relative invariants of the ruled manifold $\mathbb{P}(N_X V \oplus \mathbb{C})$.

Putting all these ingredients together, we can at last state the main result of [16].

Theorem 2.1 *Let Z be the symplectic sum of (X,V) and (Y,V) and fix a decomposition of the constraints α into α_X on the X side and α_Y on the Y side. Then the GT invariant of Z is given in terms of the relative invariants of (X,V) and (Y,V) by*

$$GT_Z(\alpha) \;=\; GT_X^V(\alpha_X) * S_V * GT_Y^V(\alpha_Y) \tag{2.3}$$

where $*$ is the convolution operation and S_V is the S-matrix defined in [16].

Several applications of this formula are described in the next two sections (see also [16] for more applications). But the full strength of the symplectic sum theorem has not yet been used.

A.-M. Li and Y. Ruan also have a sum formula [18]. Eliashberg, Givental, and Hofer are developing a general theory for invariants of symplectic manifolds glued along contact boundaries [3]. Jun Li has recently adapted our proof to the algebraic case [19].

3. Relations in $H^*(\mathcal{M}_{g,n})$

A smooth genus g curve with n marked points is stable if $2g-2+n>0$. The set of such curves, modulo diffeomorphisms, forms the moduli space $\mathcal{M}_{g,n}$. The stability condition assures that the group of diffeomorphisms acts with finite stabilizers, and so $\mathcal{M}_{g,n}$ has a natural orbifold structure. Its Deligne-Mumford compactification $\overline{\mathcal{M}}_{g,n}$ is a projective variety. Elements of $\overline{\mathcal{M}}_{g,n}$ are called stable curves; these are connected unions of smooth stable components C_i joined at d double points with a total of n marked points and Euler characteristic $\chi = 2-2g+d$. The compactification $\overline{\mathcal{M}}_{g,n}$ is also an orbifold, and in fact Looijenga proved that it has a finite degree cover which is a smooth manifold. In any event, the rational cohomology of $\overline{\mathcal{M}}_{g,n}$ satisfies Poincaré duality. Throughout this section we work only with rational coefficients.

There are several maps between moduli spaces of stable curves. First, there is a projection $\pi_i : \overline{\mathcal{M}}_{g,n+1} \to \overline{\mathcal{M}}_{g,n}$ that forgets the marked point x_i (and collapses the components that become unstable). Second, we can consider the attaching maps that build a boundary stratum in $\overline{\mathcal{M}}_{g,n}$. For each topological type of a stable curve with d nodes, with components C_i of genus g_i and n_i marked points the attaching map ξ at the d nodes takes $\sqcup_i \overline{\mathcal{M}}_{g_i,n_i}$ onto a boundary stratum of $\overline{\mathcal{M}}_{g,n}$.

We focus next on three kinds of natural classes in $H^*(\overline{\mathcal{M}}_{g,n})$ (or the Chow ring). For each i between 1 and n let $L_i \to \overline{\mathcal{M}}_{g,n}$ denote the relative cotangent bundle to the stable curve at the marked point x_i. The fiber of L_i over a point $C = (\Sigma, x_1, \ldots, x_n) \in \overline{\mathcal{M}}_{g,n}$ is the cotangent space to Σ at x_i, and its first Chern class ψ_i is called a *descendant* class. So there are n descendant classes $\psi_1, \ldots, \psi_n$, one for each marked point. Next, there are *tautological* (or Mumford-Morita-Miller) classes $\kappa_0, \kappa_1, \ldots$ obtained from powers of descendants by the formula $\kappa_a = (\pi_{n+1})_*(\psi_{n+1}^{a+1})$ for each $a \geq 0$ (where π_* denotes the push forward map in cohomology defined using the Poincaré duality). Finally, the Poincaré dual of a boundary stratum is called a *boundary* class. These three kinds of natural classes are all algebraic and even dimensional; we define their *degree* to be their complex dimension.

One natural — and difficult — problem is to describe the structure of the cohomology rings of $\mathcal{M}_{g,n}$ and $\overline{\mathcal{M}}_{g,n}$. This arises from a different perspective as well since $H^*(\mathcal{M}_{g,n})$ is also the cohomology of the mapping class group (for more details, see Tillman's I.C.M. talk). In genus zero Keel [17] determined the cohomology ring of $\overline{\mathcal{M}}_{0,n}$ in terms of generators (which are boundary classes) and relations. For higher genus far less is known about the cohomology ring.

In this section we will instead focus on finding relations in the cohomology ring. For example, in genus 0 all relations come from the "4-point relation", essentially that in the cohomology of $\overline{\mathcal{M}}_{0,4} \cong \mathbb{P}^1$ the four ψ_i classes as well as the three boundary classes are all cohomologous (all being Poincaré dual to a point). In genus 1 it is also known that ψ_1 is equal to $1/12$ of the boundary class in $\overline{\mathcal{M}}_{1,1}$. One might wonder whether in higher genus all the ψ classes come from the boundary. That turns out not to be true in genus $g \geq 2$, but in genus 2 Mumford [23] found a relation in $\overline{\mathcal{M}}_{2,1}$ expressing ψ_1^2 as a combination of boundary classes. Several years ago, Getzler [6] found a similar relation for $\psi_1\psi_2$ in $\overline{\mathcal{M}}_{2,2}$ and he conjectured that this pattern would continue in higher genus. In fact,

Theorem 3.1 *When $g \geq 1$, any product of descendant or tautological classes of degree at least g (or at least $g-1$ when $n = 0$) vanishes when restricted to $H^*(\mathcal{M}_{g,n}, \mathbb{Q})$.*

This result was proved by the author in [11]. It extends an earlier result of Looijenga [20], who proved that a product of descendant classes of degree at least $g+n-1$ vanishes in the Chow ring $A^*(\mathcal{C}_g^n)$ of the moduli space $\mathcal{C}_g^n$ of smooth genus g curves with n not necessarily distinct points.

The idea of proof of Theorem 3.1 is simple. We start with the moduli space $\mathcal{Y}_{d,g,n}$ of degree d holomorphic maps from smooth genus g curves with n marked points to S^2 which have a fixed ramification pattern over r marked points in the target. We then consider its relative stable map compactification $\overline{\mathcal{Y}}_{d,g,n}$ (closely related to the space of admissible covers [9]). The space $\overline{\mathcal{Y}}_{d,g,n}$ has an orbispace structure and it comes with two natural maps st and q that record respectively the domain and the target of the cover.

$$\begin{array}{ccc} & \overline{\mathcal{Y}}_{d,g,n} & \\ {\scriptstyle st}\swarrow & & \searrow{\scriptstyle q} \\ \overline{\mathcal{M}}_{g,n} & & \overline{\mathcal{M}}_{0,r} \end{array} \tag{3.1}$$

A simple way to get relations in the cohomology of $\overline{\mathcal{M}}_{g,n}$ is to pull back by q known relations in the cohomology of $\overline{\mathcal{M}}_{0,r}$, and then push them forward by st.

To begin with, note that the diagram above provides several other natural classes in $\overline{\mathcal{M}}_{g,n}$: for each choice of ramification pattern, $st_*\overline{\mathcal{Y}}_{d,g,n}$ defines a cycle in $\overline{\mathcal{M}}_{g,n}$. The most useful ones turn out to be the "2-point ramification cycles", for which all but at most two of the branch points are simple. Pushing forward such cycles by the attaching map of a boundary stratum gives a generalized 2-point cycle.

To prove Theorem 3.1, we choose a degree d of the cover and a 2-point ramification cycle $\overline{\mathcal{Y}}_{d,g,n}$ in such a way that the stabilization map $st : \overline{\mathcal{Y}}_{d,g,n} \to \overline{\mathcal{M}}_{g,n}$ has finite, nonzero degree. The key step is the following proposition.

Proposition 3.2 *The Poincaré dual of any degree m product of descendant and tautological classes can be written as a linear combination of generalized 2-point ramification cycles of codimension m.*

But the codimension of a 2-point ramification cycle is at most g. A simple degeneration argument proves that the cycles of codimension exactly g vanish on $\mathcal{M}_{g,n}$, thus implying Theorem 3.1.

There are three main ingredients in the proof of Proposition 3.2. First, the relative cotangent bundle to the domain is related to the pullback of the relative cotangent bundle to the target, so we can express the descendant classes in the domain via descendant classes in the target. Second, the target has genus zero and (nontrivial) products of descendants in $\overline{\mathcal{M}}_{0,r}$ are Poincaré dual to boundary cycles D. This means that we can relate a product of descendants on the domain to cycles of type st_*q^*D. Finally, a degeneration formula, which is essentially a consequence of the symplectic sum Theorem 2.1, expresses cycles of type st_*q^*D in terms of 2-point ramification cycles.

The degree g in Theorem 3.1 is the lowest degree in which some *monomial* in descendants would vanish on $\mathcal{M}_{g,n}$ (see the discussion in [10]). However, there are lower degree polynomial relations in descendent and tautological classes. For example, if we restrict our attention to the moduli space $\mathcal{M}_g$ of smooth genus g curves then the subring generated by the tautological classes is called the *tautological ring* R_g^*. Looijenga's result [20] implies that $R_g^* = 0$ for $* \geq g-1$ and Faber [4] made the following

Conjecture 3.3 *The classes $\kappa_1, \ldots, \kappa_{[g/3]}$ generate the tautological ring R_g^*.*

We refer the reader to [4] for the full conjecture.

It turns out that techniques similar to those of Theorem 3.1 produce several other sets of relations between tautological classes. One such set of relations implies that, for each $a > [g/3]$, the class κ_a can be written as polynomial in lower degree tautological classes, as required by Faber's conjecture. A detailed proof will appear in [11].

4. Further applications

There are other applications of the sum formula (2.3). One such application considered in [16] begins with the following simple observation. Given any symplectic manifold X with a codimension 2 symplectic submanifold V, we can write X as a (trivial) symplectic sum $X\#_V P_V$ where P_V is the ruled manifold $\mathbb{P}(N_X V \oplus \mathbb{C})$ and V is identified with its infinity section. We can then obtain recursive formulas for the GW invariants of X by moving constraints from one side to the other and applying the symplectic sum formula.

In [15] we used this method to obtain both (a) the Caporaso-Harris formula for the number of nodal curves in $\mathbb{P}^2$ [2], and (b) the "quasimodular form" expression for the rational enumerative invariants of the rational elliptic surface [1]. In hindsight, our proof of (a) is essentially the same as that in [2]; using the symplectic sum formula makes the proof considerably shorter and more transparent, but the key ideas are the same. Our proof of (b), however, is completely different from that of Bryan and Leung in [1].

We end with another interesting application of the Symplectic Sum Theorem 2.1. For each symplectomorphism f of a symplectic manifold X, one can form the *symplectic mapping cylinder*

$$X_f = X \times \mathbb{R} \times S^1/\mathbb{Z} \tag{4.1}$$

where the $\mathbb{Z}$ action is generated by $(x, s, \theta) \mapsto (f(x), s+1, \theta)$. In a joint paper [13] with T. H. Parker we regarded X_f as a symplectic sum and computed the Gromov invariants of the manifolds X_f and of fiber sums of the X_f with other symplectic manifolds. The result is a large set of interesting non-Kähler symplectic manifolds with computational ways of distinguishing them. In dimension four this gives a symplectic construction of the 'exotic' elliptic surfaces of Fintushel and Stern [5]. In higher dimensions it gives many examples of manifolds which are diffeomorphic but not 'equivalent' as symplectic manifolds.

More precisely, fix a symplectomorphism f of a closed symplectic manifold X, and let f_{*k} denote the induced map on $H_k(X;\mathbb{Q})$. Note that X_f fibers over the torus T^2 with fiber X. If $\det(I - f_{*1}) = \pm 1$ then there is a well-defined section class T. Our main result of [13] computes the genus one Gromov invariants of the multiples of this section class. These are the particular GW invariants that, in dimension four, C.H. Taubes related to the Seiberg-Witten invariants(see [27] and [12]).

Theorem 4.1 *If $\det(I - f_{*1}) = \pm 1$, the partial Gromov series of X_f for the section class T is given by the Lefschetz zeta function of f in the variable $t = t_T$:*

$$Gr^T(X_f) \;=\; \zeta_f(t) \;=\; \frac{\prod_{k\ odd} \det(I - t f_{*k})}{\prod_{k\ even} \det(I - t f_{*k})}.$$

When X_f is a four-manifold, a wealth of examples arise from knots. Associated to each fibered knot K in S^3 is a Riemann surface Σ and a monodromy diffeomorphism f_K of Σ. Taking $f = f_K$ gives symplectic 4-manifolds X_K of the homology type of $S^2 \times T^2$ with

$$Gr(X_K) = \frac{A_K(t_T)}{(1 - t_T)^2}$$

where $A_K(t) = \det(I - tf_{*1})$ is the Alexander polynomial of K and T is the section class.

We can elaborate on this construction by fiber summing X_f with other 4-manifolds. For example, let $E(n)$ be the simply-connected minimal elliptic surface with fiber F and holomorphic Euler characteristic n. Then $E(1)$ is the rational elliptic surface and $K3 = E(2)$. Forming the fiber sum of X_K with $E(n)$ along the tori $T = F$, we obtain a symplectic manifold

$$E(n, K) \;=\; E(n)\#_{F=T}X_K.$$

homeomorphic to $E(n)$. In fact, for fibered knots K, K' of the same genus there is a homeomorphism between $E(n, K)$ and $E(n, K')$ preserving the periods of ω and the canonical class κ. For $n > 1$ we can compute the full (not just partial) Gromov series.

Proposition 4.2 *For $n \geq 2$, the Gromov and Seiberg-Witten series of $E(n, K)$ are*

$$Gr(E(n, K)) \;=\; SW(E(n, K)) \;=\; A_K(t_F)\ (1 - t_F)^{n-2}. \tag{4.2}$$

Thus fibered knots with distinct Alexander polynomials give rise to symplectic manifolds $E(n, K)$ which are homeomorphic but not diffeomorphic. In particular, there are infinitely many distinct symplectic 4-manifolds homeomorphic to $E(n)$. Fintushel and Stern [5] have independently shown how (4.2) follows from knot theory and results in Seiberg-Witten theory.

References

[1] J. Bryan and N.-C. Leung, *The enumerative geometry of $K3$ surfaces and modular forms*, J. Amer. Math. Soc. **13** (2000), 371–410.

[2] L. Caporaso and J. Harris, *Counting plane curves in any genus*, Invent. Math. **131** (1998), 345–392.

[3] Y. Eliashberg, A. Givental and H. Hofer, *Introduction to Symplectic Field Theory*, GAFA 2000 (Tel Aviv, 1999), Geom. Funct. Anal. 2000, Special Volume, Part II, 560–673.

[4] C. Faber, *A conjectural description of the tautological ring of the moduli space of curves*, Moduli of curves and abelian varieties, 109–129, Aspects Math., E33, Vieweg, Braunschweig, 1999.

[5] R. Fintushel and R. Stern, *Knots, Links and 4-Manifolds*, Invent. Math. **134** (1998), 363–400.

[6] E. Getzler, *Topological recursion relations in genus 2*, Integrable systems and algebraic geometry (Kobe/Kyoto, 1997), 73–106, World Sci. Publishing, River Edge, NJ, 1998.

[7] R. Gompf, *A new construction of symplectic manifolds*, Annals of Math., **142** (1995), 527–595.

[8] M. Gromov, *Pseudo holomorphic curves in symplectic manifolds*, Invent. Math. **82** (1985), 307–347.

[9] J. Harris, I. Morrison, *Moduli of curves*, Graduate Texts in Math, vol 187, Springer-Verlag, 1998.
[10] E. Ionel, *Topological recursive relations in* $H^{2g}(\mathcal{M}_{g,n})$, to appear in Invent. Math.
[11] E. Ionel, *On relations in the tautological ring of* $\mathcal{M}_g$, in preparation.
[12] E. Ionel and T. H. Parker, *The Gromov invariants of Ruan-Tian and Taubes*, Math. Res. Lett. **4** (1997), 521–532.
[13] E. Ionel and T. H. Parker, *Gromov Invariants and Symplectic Maps*, Math. Annalen, **314**, 127–158 (1999).
[14] E. Ionel and T. H. Parker, *Gromov-Witten Invariants of Symplectic Sums*, announcement, Math. Res. Lett., **5**(1998), 563–576.
[15] E. Ionel and T. H. Parker, *Relative Gromov-Witten Invariants*, to appear in Annals of Math.
[16] E. Ionel and T. H. Parker, *The Symplectic Sum Formula for Gromov-Witten Invariants*, preprint, math.SG/0010217.
[17] S. Keel, *Intersection theory of moduli space of stable n-pointed curves of genus zero*, Trans. Amer. Math. Soc. **330**(1992), 545–574.
[18] A.-M. Li, Y. Ruan, *Symplectic surgery and Gromov-Witten invariants of Calabi-Yau 3-folds*, Invent. Math. **145** (2001), 151–218.
[19] Jun Li, *A Degeneration formula of GW-invariants*, preprint, math.AG/0110113.
[20] E. Looijenga, *On the tautological ring of* $\mathcal{M}_g$, Invent. Math. **121**(1995), 411–419.
[21] J. McCarthy and J.Wolfson, *Symplectic Normal Connect Sum*, Topology, **33** (1994) 729–764.
[22] D. McDuff and D. Salamon, J-holomorphic curves and quantum cohomology, A.M.S., Providence, R.I., 1994.
[23] D. Mumford, *Towards an enumerative geometry of the moduli space of curves* in *Arithmetic and geometry II* (ed. M. Artin and J. Tate), Progress in Math, vol 36, Birkhäuser, Basel, 1983.
[24] T. H. Parker and J. Wolfson, *Pseudo-holomorphic maps and bubble trees*, Jour. Geometric Analysis, **3** (1993) 63–98.
[25] Y. Ruan and G. Tian, *A mathematical theory of quantum cohomology*, J. Differential Geom. **42** (1995), 259–367.
[26] Y. Ruan and G. Tian, *Higher genus symplectic invariants and sigma models coupled with gravity*, Invent. Math. **130** (1997), 455–516.
[27] C. H. Taubes, *Counting pseudo-holomorphic curves in dimension four*, J. Diff. Geom. **44** (1996), 818–893.

ICM 2002 · Vol. II · 437–446

Knots, von Neumann Signatures, and Grope Cobordism*

Peter Teichner†

Abstract

We explain new developments in classical knot theory in 3 and 4 dimensions, i.e. we study knots in 3-space, up to isotopy as well as up to concordance. In dimension 3 we give a geometric interpretation of the Kontsevich integral (joint with Jim Conant), and in dimension 4 we introduce new concordance invariants using von Neumann signatures (joint with Tim Cochran and Kent Orr). The common geometric feature of our results is the notion of a grope cobordism.

2000 Mathematics Subject Classification: 57M25, 57N70, 46L89.
Keywords and Phrases: Knot, Signature, von Neumann algebra, Concordance, Kontsevich integral, Grope.

1. Introduction

A lot of fascinating mathematics has been created when successful tools are transferred from one research area to another. We shall describe two instances of such transfers, both into knot theory. The first transfer realizes commutator calculus of group theory by *embedded versions* in 3- and 4-space, and produces many interesting geometric equivalence relations on knots, called *grope cobordism* in 3-space and *grope concordance* in 4-space. It turns out that in 3-space these new equivalence relations give a geometric interpretation (Theorem 2) of Vassiliev's finite type invariants [21] and that the Kontsevich integral [17] calculates the new theory over $\mathbb{Q}$ (Theorem 3).

In 4-space the new equivalence relations factor naturally through knot concordance, and in fact they organize all known concordance invariants in a wonderful manner (Theorem 5). They also point the way to new concordance invariants (Theorem 6) and these are constructed using a second transfer, from the spectral theory of self-adjoint operators and von Neumann's continuous dimension [20].

*Partially supported by an NSF-grant and the Max-Planck Gesellschaft.
†University of California in San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0112, USA. E-mail: teichner@math.ucsd.edu

1.1. A geometric interpretation of group commutators

To explain the first transfer into knot theory, recall that every knot bounds a Seifert surface (embedded in 3-space), but only the trivial knot bounds an embedded disk. Thus all of knot theory is created by the difference between a surface and a disk. The new idea is to filter this difference by introducing a concept into knot theory which is the analogue of iterated commutators in group theory. Commutators arise because a continuous map $\phi : S^1 \to X$ extends to a map of a surface if and only if ϕ represents a commutator in the fundamental group $\pi_1 X$. Iterated commutators can similarly be expressed by gluing together several surfaces. Namely, there are certain finite 2-complexes (built out of iterated surface stages) called *gropes* by Cannon [1], with the following defining property: $\phi : S^1 \to X$ represents an element in the k-th term of the *lower central series* of $\pi_1 X$ if and only if it extends to a continuous map of a *grope of class* k. Similarly, there are *symmetric gropes* which geometrically implement the derived series of $\pi_1 X$, see Figures 2. and 3.

Gropes, therefore, are not quite manifolds but the singularities that arise are of a very simple type, so that these 2-complexes are in some sense the next easiest thing after surfaces. Two sentences on the history of the use of gropes in mathematics are in place, compare [11, Sec.2.11]. Their inventor Stan'ko worked in high-dimensional topology, and so did Edwards and Cannon who developed gropes further. Bob Edwards suggested their relevance for topological 4-manifolds, where they were used extensively, see [11] or [12]. It is this application that seems to have created a certain "Angst" of studying gropes, so we should point out that the only really difficult part in that application is the use of *infinite constructions*, i.e. when the class of the grope goes to infinity.

One purpose of this note is to convince the reader that (finite) gropes are a very simple, and extremely powerful tool in low-dimensional topology. The point is that once one can describe iterated commutators in $\pi_1 X$ by maps of gropes, one might as well study *embedded* gropes in 3-space (respectively 4-space) in order to organize knots up to isotopy (respectively up to concordance). In Section 2. we shall explain joint work with Jim Conant on how gropes embedded in 3-space lead to a geometric interpretation of Vassiliev's knot invariants [21] and of the Kontsevich integral [17].

1.2. von Neumann signatures and knot concordance

In Section 3. we study symmetric gropes embedded in 4-space, and explain how they lead to a geometric framework for all known knot concordance invariants and beyond. More precisely, we explain our joint work with Tim Cochran and Kent Orr [6], where we define new concordance invariants by inductively constructing representations into certain solvable groups G, and associating a hermitian form over the group ring $\mathbb{Z}G$ to the knot K, which is derived from the intersection form of a certain 4-manifold with fundamental group G and whose boundary is obtained by 0-surgery on K. This intersection form represents an element in the Cappell-Shaneson Γ-group [2] of $\mathbb{Z}G$ and we detect it via the second transfer from a different area of mathematics: The standard way to detect elements in Witt groups like the

Γ-group above is to construct unitary representations of G, and then consider the corresponding (twisted) signature of the resulting hermitian form over $\mathbb{C}$. It turns out that the solvable groups G we construct do not have any interesting finite dimensional representations, basically because they are "too big" (e.g. not finitely generated), a property that is intrinsic to the groups G in question because they are "universal solvable" in the sense that many 4-manifold groups (with the right boundary) must map to G, extending the given map of the knot group.

However, every group G has a fundamental unitary representation given by $\ell^2 G$, the Hilbert space of square summable sequences of group elements with complex coefficients. The resulting (weak) completion of $\mathbb{C}G$ is the *group von Neumann algebra* $\mathcal{N}G$. It is of type II_1 because the map $\sum a_g g \mapsto a_1$ extends from $\mathbb{C}G$ to give a finite faithful trace on $\mathcal{N}G$.

The punchline is that hermitian forms over the completion $\mathcal{N}G$ are much easier to understand than over $\mathbb{C}G$ because they are diagonalizable (by functional calculus of self-adjoint operators). Here one really uses the von Neumann algebra, rather than the C^*-algebra completion of $\mathbb{C}G$ because the functional calculus must be applied to the characteristic functions of the positive (respectively negative) real numbers, which are bounded but *not* continuous.

The subspace on which the hermitian form is positive (respectively negative) definite has a *continuous dimension*, which is the positive real number given by the trace of the projection onto that subspace. As a consequence, one can associate to every hermitian form over $\mathcal{N}G$ a real valued invariant, the *von Neumann signature*. In [6] we use this invariant to construct our new knot concordance invariants, and a survey of this work can be found in Section 3. It is not only related to embedded gropes in 4-space but also to the existence of towers of *Whitney disks* in 4-space. Unfortunately, we won't be able to explain this aspect of the theory, but see [6, Thm.8.12].

1.3. Noncommutative Alexander modules

In Section 3. we shall hint at how the interesting representations to our solvable groups are obtained. But it is well worth pointing out that the methods developed for studying knot concordance have much simpler counterparts in 3-space, i.e. if one is only interested in isotopy invariants.

A typical list of knot invariants that might find its way into a text book or survey talk on *classical* knot theory, would contain the Alexander polynomial, (twisted) signatures, (twisted) Arf invariants, and maybe knot determinants. It turns out that all of these invariants can be computed from the homology of the infinite cyclic covering of the knot complement, and are in this sense "commutative" invariants.

Instead of the maximal abelian quotient one can use other solvable quotient groups of the knot group to obtain "noncommutative" knot invariants. The canonical candidates are the quotient groups G_n of the derived series $G^{(n)}$ of the knot group (compare Section 3. for the definition). One can thus define the higher order

Alexander modules of a knot K to be the $\mathbb{Z}G_{n+1}$-modules

$$\mathcal{A}_n(K) := H_1(S^3 \setminus K; \mathbb{Z}G_{n+1}).$$

The indexing is chosen so that $\mathcal{A}_0$ is the classical Alexander module. For $n \geq 1$ these modules are best studied by introducing further algebraic tools as follows. By a result of Strebel the groups G_n are torsionfree. Therefore, the group ring $\mathbb{Z}G_n$ satisfies the Ore condition and has a well defined (skew) quotient field. This field is in fact the quotient field of a (skew) polynomial ring $\mathbb{K}_n[t^{\pm 1}]$, with $\mathbb{K}_n$ the quotient field of $\mathbb{Z}[G^{(1)}/G^{(n)}]$ and $G_1 = \langle t \rangle \cong \mathbb{Z}$. Thus one is exactly in the context of [6, Sec.2] and one can define explicit noncommutative isotopy invariants of knots. For example, let $d_n(K)$ be the dimension (over the field $\mathbb{K}_{n+1}$) of the *rational* Alexander module

$$\mathcal{A}_n(K) \otimes_{\mathbb{Z}G_{n+1}} \mathbb{K}_{n+1}[t^{\pm 1}].$$

It is shown in [6, Prop.2.11] that these dimensions are finite with the degree of the usual Alexander polynomial being $d_0(K)$. Moreover, Cochran [5] has proven the following non-triviality result for these dimensions.

Theorem *If K is a nontrivial knot then for $n \geq 1$ one has*

$$d_0(K) \leq d_1(K) + 1 \leq d_2(K) + 1 \leq \cdots \leq d_n(K) + 1 \leq 2 \cdot \textit{genus of } K.$$

Moreover, there are examples where these numbers are strictly increasing up to any given n.

Corollary *If one of the inequalities in the above theorem is strict then K is not fibered. Furthermore, 0-surgery on K cross the circle is not a symplectic 4-manifold.*

The first statement is clear: For fibered knots the degree of the Alexander polynomial $d_0(K)$ equals twice the genus of the knot K. The second statement follows from a result of Kronheimer [18] who showed that this equality also holds if the above 4-manifold is symplectic.

Recently, Harvey [16] has studied similar invariants for arbitrary 3-manifolds and has proven generalizations of the above results: There are lower bounds for the Thurston norm of a homology class, analogous to $d_i(K)$, that are better than McMullen's lower bound, which is the analogy of $d_0(K)$. As a consequence, she gets new algebraic obstructions to a 4-manifold of the form $M^3 \times S^1$ admitting a symplectic structure.

Just like in the classical case $n = 0$, there is more structure on the rational Alexander modules. By [6, Thm.2.13] there are higher order *Blanchfield forms* which are hermitian and non-singular in an appropriate sense, compare [5, Prop.12.2]. It would be very interesting to know whether the n-th order Blanchfield form determines the von Neumann η-invariant associated to the G_{n+1}-cover. So far, these η-invariants are very mysterious real numbers canonically associated to a knot.

Only in the bottom case $n = 0$ do we understand this η-invariant well: The L^2-index theorem implies that the von Neumann η-invariant corresponding to the $\mathbb{Z}$-cover is the von Neumann signature of a certain 4-manifold with fundamental

group $\mathbb{Z}$. Moreover, this signature is the integral, over the circle, of all (Levine-Tristram) twisted signatures of the knot [7, Prop.5.1] (and is thus a concordance invariant). For $n \geq 1$ there is in general no such 4-manifold available and the corresponding η-invariants are not concordance invariants.

2. Grope cobordism in 3-space

We first give a more precise treatment of the first transfer from group theory to knot theory hinted at in the introduction. Recall that the fundamental group consists of continuous maps of the circle S^1 into some target space X, modulo homotopy (i.e. 1-parameter families of continuous maps). Quite analogously, classical knot theory studies smooth *embeddings* of a circle into S^3, modulo isotopy (i.e. 1-parameter families of embeddings). To explain the transfer, we recall that a continuous map $\phi : S^1 \to X$ represents the trivial element in the fundamental group $\pi_1 X$ if and only if it extends to a map of the disk, $\tilde{\phi} : D^2 \to X$. Moreover, ϕ represents a commutator in $\pi_1 X$ if and only if it extends to a map of a surface (i.e. of a compact oriented 2-manifold with boundary S^1). The first statement has a straightforward analogy in knot theory: $K : S^1 \hookrightarrow S^3$ is trivial if and only if it extends to an embedding of the disk into S^3. However, every knot "is a commutator" in the sense that it bounds a *Seifert surface*, i.e. an embedded surface in S^3.

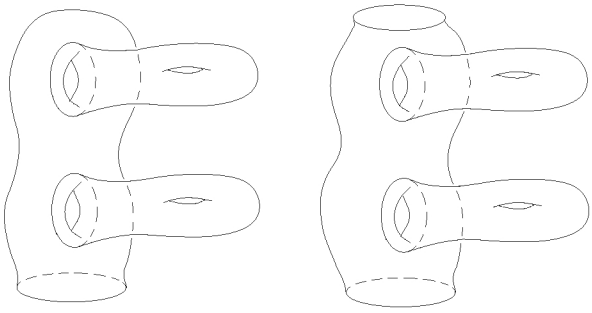


Figure 1: Gropes of class 3, with one respectively two boundary circles

Recall from the introduction that gropes are finite 2-complexes defined by the following property: $\phi : S^1 \to X$ represents an element in the k-th term $\pi_1 X_k$ of the lower central series of $\pi_1 X$ if and only if it extends to a continuous map of a grope of class k. Here G_k is defined inductively for a group G by the iterated commutators

$$G_2 := [G, G] \text{ and } G_k := [G, G_{k-1}] \text{ for } k > 2.$$

Accordingly, a grope of class 2 is just a surface, and one can obtain a grope of class k by attaching gropes of class $(k-1)$ to g disjointly embedded curves in the *bottom* surface. Here g is the genus of the bottom surface and the curves are assumed to span one half of its homology. This gives gropes of class k with one boundary circle as on the left of Figure 2. It's not the most general way to get gropes because of

re-bracketing issues, and we refer to [9, Sec.2.1] for details. The boundary of a grope is by definition just the boundary of the bottom surface, compare Figure 2.

Definition 1 *Two (smooth oriented) knots in S^3 are* grope cobordant *of class k, if there is an embedded grope of class k in S^3 (the* grope cobordism*) such that its boundary consists exactly of the given knots.*

An *embedding* of a grope is best defined via the obvious 3-dimensional local model. Since every grope has a 1-dimensional spine, embedded gropes can then be isotoped into the neighborhood of a 1-complex. As a consequence, embedded gropes abound in 3-space! It is important to point out that if two knots K_i cobounds a grope then K_1 and K_2 might very well be linked in a nontrivial way. Thus it is a much stronger condition on K to assume that it is the boundary of an embedded grope than to say that it cobounds a grope with the unknot. For example, if K bounds an embedded grope of class 3 in S^3 then the Alexander polynomial vanishes. Together with Stavros Garoufalidis, we recently showed [13] that the 2-loop term of the Kontsevich integral detects many counterexamples to the converse of this statement.

In joint work with Jim Conant [9], we show that grope cobordism defines equivalence relations on knots, one for every class $k \in \mathbb{N}$. Moreover, Theorem 2 below implies that the resulting quotients are in fact finitely generated abelian groups (under the connected sum operation). For the smallest values $k = 2, 3, 4$ and 5, these groups are isomorphic to

$$\{0\}, \mathbb{Z}/2, \mathbb{Z} \text{ and } \mathbb{Z} \times \mathbb{Z}/2$$

and they are detected by the first two Vassiliev invariants [10, Thm.4.2].

The following theorem is formulated in terms of *clasper surgery* which was introduced independently by Habiro [15] and Goussarov [14], as a geometric answer to finite type invariants a lá Vassiliev [21]. We cannot explain the definitions here but see [9, Thm.1 and 3]. We should say that the notion of a *capped* grope is well known in 4 dimensions, see [11, Sec.2]. In our context, it means that all circles at the "tips" of the grope bound disjointly embedded disks in 3-space which are only allowed to intersect the bottom surface of the grope.

Theorem 2 *Two knots K_0 and K_1 are grope cobordant of class k if and only if K_1 can be obtained from K_0 by a finite sequence of clasper surgeries of* grope degree k *(as defined below).*

Moreover, two knots are capped *grope cobordant of class k if and only if they have the same finite type invariants of Vassiliev degree $< k$.*

As a consequence of this result, the invariants associated to grope cobordism are highly nontrivial as well as manageable. For example, we prove the following result in [10, Thm.1.1]:

Theorem 3 *The (logarithm of the) Kontsevich integral (with values in $\mathcal{B}^g_{<k}$), graded by the new* grope degree k*, is an obstruction to finding a grope cobordism of class k between two knots. Moreover, this invariant is rationally faithful and surjective.*

Here $\mathcal{B}^g_{<k}$ is one of the usual algebras of Feynman diagrams known from the theory of finite type invariants, but graded by the *grope degree*. More precisely, $\mathcal{B}^g_{<k}$ is the $\mathbb{Q}$-vector space generated by connected uni-trivalent graphs of grope degree $i, 1 < i < k$, with at least one univalent vertex and a cyclic ordering at each trivalent vertex. The relations are the usual IHX and AS relations. The grope degree is the Vassiliev degree (i.e. half the number of vertices) *plus* the first Betti number of the graph. Observe that both relations preserve this new degree.

Read backwards, our results give an interpretation of the Kontsevich integral in terms of the geometrically defined equivalence relations of grope cobordism.

3. Grope concordance

We now turn to the 4-dimensional aspects of the theory. It may look like the end of the story to realize that any knot with trivial Arf invariant bounds a grope of arbitrary big class embedded in D^4, [10, Prop.3.8]. However, group theory has more to offer than the lower central series. Recall that the *derived series* of a group G is defined inductively by the iterated commutators

$$G^{(1)} := [G, G] \text{ and } G^{(h)} := [G^{(h-1)}, G^{(h-1)}] \text{ for } h > 1.$$

Accordingly, we may define *symmetric gropes* with their complexity now measured by *height*, satisfying the following defining property: A continuous map $\phi : S^1 \to X$ represents an element in $\pi_1 X^{(h)}$ if and only if it extends to a continuous map of a symmetric grope of height h. Thus a symmetric grope of height 1 is just a surface, and a symmetric grope of height h is obtained from a bottom surface by attaching symmetric gropes of height $(h-1)$ to a *full* symplectic basis of curves. This defines symmetric gropes of height h with one boundary circle as in Figure 3.

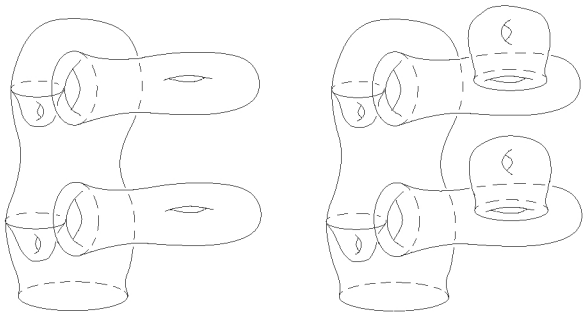


Figure 2: Symmetric gropes of height 2 and 2.5

Note that a symmetric grope of height h is also a grope of class 2^h, just like in group theory. But conversely, not every grope is symmetric. It should also be clear from Figure 3. how one defines symmetric gropes with half-integer height (even though there is no group theoretic analogue).

In the following definition we attempt to distinguish the terms "cobordant" and "concordant" in the sense that the latter refers to 4 dimensions, whereas the former was used in dimension 3, see Definition 1. Historically, these terms were used interchangeably, but we hope not to create any confusion with our new distinction.

Definition 4 *Two oriented knots in* S^3 *are* grope concordant *of height* $h \in \frac{1}{2}\mathbb{N}$, *if there is an embedded symmetric grope of height* h *in* $S^3 \times [0,1]$ *such that its boundary consists exactly of the given knots* $K_i : S^1 \hookrightarrow S^3 \times \{i\}$.

Observe that since an annulus is a symmetric grope of arbitrary height we indeed get a filtration of the *knot concordance group*. This group is defined by identifying two knots which cobound an embedded annulus in $S^3 \times [0,1]$, where there are two theories depending on whether the embedding is smooth or just topological (and locally flat). For grope concordance the smaller topological knot concordance group is the more natural setting: a locally flat topological embedding of a grope (defined by the obvious local model at the singular points) can be perturbed to become smooth. This perturbation might introduce many self-intersection points in the surface stages of the grope. However, these new singularities are arbitrarily small and thus they can be removed at the expense of increasing the genus of the surface stage in question but without changing the height of the grope.

In joint work with Tim Cochran and Kent Orr [6], we showed that all known knot concordance invariants fit beautifully into the scheme of grope concordance! In particular, all known invariants turned out to already be invariants of grope concordance of height 3.5:

Theorem 5 *Consider two knots* K_i *in* S^3. *Then*

1. K_i *have the same Arf invariant if and only if they are grope concordant of height* 1.5 *(or class 3).*
2. K_i *are* algebraically concordant *in the sense of Levine [19] (i.e. all twisted signatures and twisted Arf invariants agree) if and only if they are grope concordant of height* 2.5.
3. *If* K_i *are grope concordant of height* 3.5 *then they have the same Casson-Gordon invariants [3].*

The third statement includes the generalizations of Casson-Gordon invariants by Gilmer, Kirk-Livingston, and Letsche. In [6] we prove an even stronger version of the third part of Theorem 5. Namely, we give a weaker condition for a knot K to have vanishing Casson-Gordon invariants: it suffices that K is (1.5)-*solvable.* All the obstruction theory in [6] is based on the definition of (h)-*solvable* knots, $h \in \frac{1}{2}\mathbb{N}$, which we shall not give here. Suffice it to say that this definition is closer to the algebraic topology of 4-manifolds than grope concordance. In [6, Thm.8.11] we show that a knot which bounds an embedded grope of height $(h+2)$ in D^4 is (h)-solvable.

It should come as no surprise that the invariants which detect grope concordance have to do with solvable quotients of the knot group. In fact, the above invariants are all obtained by studying the Witt class of the intersection form of a

certain 4-manifold M^4 whose boundary is obtained by 0-surgery on the knot. The different cases are distinguished by the fundamental group $\pi_1 M$, namely

1. $\pi_1 M$ is trivial for the Arf invariant,
2. $\pi_1 M$ is infinite cyclic for algebraic concordance, and
3. $\pi_1 M$ is a dihedral group for Casson-Gordon invariants.

So the previously known concordance invariants stopped at solvable groups which are extensions of abelian by abelian groups. To proceed further in the understanding of grope (and knot) concordance, one must be able to handle more complicated solvable groups. A program for that purpose was developed in [6] by giving an elaborate boot strap argument to construct inductively representations of knot groups into certain *universal solvable* groups. On the way, we introduced Blanchfield duality pairings in solvable covers of the knot complement by using noncommutative localizations of the group rings in question.

The main idea of the boot strap is that a particular choice of "vanishing" of the previous invariant *defines* the map into the next solvable group (and hence the next invariant). In terms of gropes this can be expressed quite nicely as follows: pick a grope concordance of height $h \in \mathbb{N}$ and use it to construct a certain 4-manifold whose intersection form gives an obstruction to being able to extend that grope to height $h.5$. There is an obvious technical problem in such an approach, already present in [3]: to show that there is no grope concordance of height $h.5$, one needs to prove non-triviality of the obstruction *for all* possible gropes of height h. One way around this problem is to construct examples where the grope concordances of small height are in some sense unique. This was done successfully in [6] for the level above Casson-Gordon invariants, and in [7] we even obtain the following infinite generation result. Let $\mathcal{G}_h$ be the graded quotient groups of knots, grope concordant of height h to the unknot, modulo grope concordance of height $h.5$. Then the results of Levine and Casson-Gordon show that $\mathcal{G}_2$ and $\mathcal{G}_3$ are not finitely generated.

Theorem 6 *$\mathcal{G}_4$ is not finitely generated.*

The easiest example of a non-slice knot with vanishing Casson-Gordon invariants is given in [6, Fig.6.5]. As explained in the introduction, the last step in the proof of Theorem 6 is to show that the intersection form of the 4-manifold in question is nontrivial in a certain Witt group. Our new tool is the von Neumann signature which has the additional bonus that it takes values in $\mathbb{R}$, which is not finitely generated as an abelian group. This fact makes the above result tractable. We cannot review any aspect of the von Neumann signature here, but see [6, Sec.5].

Last but not least, it should be mentioned that we now know that for every $h \in \mathbb{N}$ the groups $\mathcal{G}_h$ are nontrivial. This work in progress [8] uses as the main additional input the Cheeger-Gromov estimate for von Neumann η-invariants [4] in order to get around the technical problem mentioned above. It is very likely that non of the groups $\mathcal{G}_h, h \in \mathbb{N}, h \geq 2$, are finitely generated.

References

[1] J. W. Cannon, *The recognition problem: what is a topological manifold?* Bull. AMS 84 (1978) 832–866.

[2] S. Cappell and J. L. Shaneson, *The codimension two placement problem and homology equivalent manifolds*, Annals of Math. 99 (1974) 227–348.

[3] A. J. Casson and C. McA. Gordon, *Cobordism of classical knots*, Orsay notes 1975, published in *A la recherche de la topologie perdue*, ed. Guilllou and Marin, Progress in Math. 1986.

[4] J. Cheeger and M. Gromov, *Bounds on the von Neumann dimension of L^2-cohomology and the Gauss-Bonnet Theorem for open manifolds*, J. Diff. Geometry 21 (1985) 1–34.

[5] T. Cochran, *Noncommutative Knot theory*, math.GT/02.

[6] T. Cochran, K. Orr and P. Teichner, *Knot concordance, Whitney Towers, and L^2-signatures*, math.GT/9908117, to appear in the Annals of Math.

[7] T. Cochran, K. Orr and P. Teichner, *Structure in the classical knot concordance group*, math.GT/0206059.

[8] T. Cochran and P. Teichner, *Von Neumann η-invariants and grope concordance*, in preparation.

[9] J. Conant and P. Teichner, *Grope cobordism of classical knots*, math.GT/0012118.

[10] J. Conant and P. Teichner, *Grope cobordism and Feynman diagrams*, math.GT/02.

[11] M. Freedman and F. Quinn, *The topology of 4-manifolds*, Princeton Math. Series 39, Princeton, NJ, 1990.

[12] M. Freedman and P. Teichner, *4-Manifold topology I+II*, Inventiones Math. 122 (1995) 509–529 and 531–557.

[13] S. Garoufalidis and P. Teichner, *On Knots with trivial Alexander polynomial*, math.GT/0206023.

[14] M. Goussarov, *Knotted graphs and a geometrical technique of n-equivalence*, St. Petersburg Math. J. 12–4 (2001).

[15] K. Habiro, *Claspers and finite type invariants of links*, Geometry and Topology, vol 4 (2000) 1–83.

[16] S. Harvey, *Higher Order Polynomial Invariants of 3-manifolds giving lower bound for the Thurston Norm*, preprint, Rice University 2001.

[17] M. Kontsevich, *Vassiliev's knot invariants*, Adv. in Sov. Math. 16(2) (1993) 137–150.

[18] P. Kronheimer, *Minimal genus in $S^1 \times M^3$*, Inventiones Math. 134 (1998) 363–400.

[19] J. P. Levine, *Knot cobordism groups in codimension two*, Comm. Math. Helv. 44 (1969) 229–244.

[20] F.J. Murray and J. von Neumann, *On Rings of Operators*, Annals of Math. 37 (1936) 116–229.

[21] V. A. Vassiliev, *Cohomology of Knot Spaces*, Theory of Singularities and its Applications, ed. V. I. Arnold, AMS, Providence 1990.

ICM 2002 · Vol. II · 447–456

Strings and the Stable Cohomology of Mapping Class Groups

Ulrike Tillmann*

2000 Mathematics Subject Classification: 57R20, 55P47, 32G15, 81T40.
Keywords and Phrases: Elliptic cohomology, Cohomology of moduli spaces, Infinite loop spaces, Cobordism theory.

1. Introduction

Twenty years ago, Mumford initiated the systematic study of the cohomology ring of moduli spaces of Riemann surfaces. Around the same time, Harer proved that the homology of the mapping class groups of oriented surfaces is independent of the genus in low degrees, increasing with the genus. The (co)homology of mapping class groups thus stabelizes. At least rationally, the mapping class groups have the same (co)homology as the corresponding moduli spaces. This prompted Mumford to conjecture that the stable rational cohomology of moduli spaces is generated by certain tautological classes that he defines. Much of the recent interest in this subject is motivated by mathematical physics and, in particular, by string theory. The study of the category of strings led to the discovery of an infinite loop space, the cohomology of which is the stable cohomology of the mapping class groups. We explain here a homotopy theoretic approach to Mumford's conjecture based on this fact. As byproducts infinite families of torsion classes in the stable cohomology are detected, and the divisibility of the tautological classes is determined. An analysis of the category of strings in a background space leads to the formulization of a parametrized version of Mumford's conjecture.

The paper is chiefly a summary of the author's work and her collaboration with Ib Madsen. Earlier this year Madsen and Weiss announced a solution of Mumford's conjecture. We touch on some of the consequences and the ideas behind this most exciting new developement.

*Math. Inst., 24–29 St. Giles, Oxford OX1 3LB, UK. E-mail: tillmann@maths.ox.ac.uk

2. Mumford's conjecture

Let $F^s_{g,n}$ be an oriented, connected surface of genus g with s marked points and n boundary components. Let $\mathrm{Diff}(F^s_{g,n})$ be its group of orientaton preserving diffeomorphisms that fix the n boundary components pointwise and permute the s marked points. By [2], for genus at least 2, $\mathrm{Diff}(F^s_{g,n})$ is homotopic to its group of components, the mapping class group $\Gamma^s_{g,n}$. Furthermore, if the surface has boundary, $\Gamma^s_{g,n}$ acts freely on Teichmüller space. Hence,

$$B\,\mathrm{Diff}(F^s_{g,n}) \simeq B\Gamma^s_{g,n} \simeq \mathcal{M}^s_{g,n} \quad \text{for } n \geq 1, g \geq 2,$$

where $\mathcal{M}^s_{g,n}$ denotes the moduli space of Riemann surfaces appropriately marked. When $n = 0$, the action of the mapping class group on Teichmüller space has finite stabilizer groups and the latter is only a rational equivalence.

We recall Harer's homology stability theorem [4] which plays an important role through out the paper.

Harer Stability Theorem 2.1. *$H_* B\Gamma^s_{g,n}$ is independent of g and n in degrees $3* < g - 1$.*

Ivanov [5], [6] improved the stability range to $2* < g - 1$ and proved a version with twisted coefficients. Glueing a torus with two boundary components to a surface $F_{g,1}$ induces a homomorphism $\Gamma_{g,1} \to \Gamma_{g+1,1}$. Let $\Gamma_\infty := \lim_{g\to\infty} \Gamma_{g,1}$ be the stable mapping class group.

Mumford [12] introduced certain tautological classes in the cohomology of the moduli spaces $\mathcal{M}_g$. Topological analogues were studied by Miller [10] and Morita [11]: Let E be the universal F-bundle over $B\,\mathrm{Diff}(F)$, and let T^vE be its vertical tangent bundle with Euler class $e \in H^2E$. Define

$$\kappa_i := \int_F e^{i+1} \quad \in H^{2i} B\,\mathrm{Diff}(F).$$

Here $\int_F$ denotes "integration over the fiber" - the Gysin map. Miller and Morita showed independently that the rational cohomology of the stable mapping class group contains the polynomial ring on the κ_i.

Mumford Conjecture 2.2. $H^*(B\Gamma_\infty; \mathbb{Q}) \simeq \mathbb{Q}[\kappa_1, \kappa_2, \dots]$.

2.1. Remark.

The stable cohomology of the decorated mapping class groups is known modulo $H^*B\Gamma_\infty$ because of decoupling [1]. For example, let $\Gamma^s_\infty := \lim \Gamma^s_{g,1}$. The following is a consequence of Theorem 2.1.

Proposition 2.3. $(B\Gamma^s_\infty)^+ \simeq B\Gamma^+_\infty \times B(\Sigma_s \wr S^1)^+$.

Here Y^+ denotes Quillen's plus-construction on Y with respect to the maximal perfect subgroup of the fundamental group. It is important to note that **the plus construction does not change the (co)homology**. In particular,

$$H^*B\Gamma_\infty = H^*B\Gamma^+_\infty.$$

3. String category

The category underlying the quantum mechanics of a state space X is the path category $\mathcal{P}X$. Its objects are particles represented by points in X. As time evolves a particle sweeps out a path. Thus a morphism between particles a and b is a continuous path in X starting at a and ending at b. Concatenation of paths defines the composition in the category.

$$\mathcal{P}X = \begin{cases} \text{objects} : & a, b, \cdots \in X, \\ \text{morphisms} : & \coprod_{t>0} \text{map}([0,t], X). \end{cases}$$

In string theory, the point objects are replaced by closed loops in X. As time evolves these strings sweep out a surface. Thus the space of morphisms from one string to another is now described by a continuous map from an oriented surface F to X. The parametrization of the path should be immaterial. To reflect this, take homotopy orbits under the action of $\text{Diff}(F)$.[1] Composition is given by concatenation of paths, i.e. by glueing of surfaces along outgoing and incoming boundary components.

To be more precise, let $LX = \text{map}(S^1, X)$ denote the free loop space on X. A cobordism F is a finite union

$$F_{g_1, n_1+m_1} \cup \cdots \cup F_{g_k, n_k+m_k}$$

where $n = \Sigma_i n_i$ boundary components are considered incoming and $m = \Sigma_i m_i$ outgoing. For technical reasons we will assume $m_i > 0$. The category of strings in X is then

$$\mathcal{S}X = \begin{cases} \text{objects} : & \alpha, \beta, \cdots \in \coprod_{n \geq 0} (LX)^n, \\ \text{morphisms} : & \coprod_F E\,\text{Diff}(F) \times_{\text{Diff}(F)} \text{map}(F, X). \end{cases}$$

The disjoint union is taken over all cobordisms F, one for each topological type.

3.1. Elliptic elements.

The category $\mathcal{S}X$ was first introduced by Segal [14]. A functor from the path category $\mathcal{P}X$ to the category of n-dimensional vector spaces and their isomorphisms defines a vector bundle on X with connection. In particular, it defines an element in the K-theory of X. A functor from $\mathcal{S}X$ to an appropriate (infinite dimensional) vector space category is also referred to as a gerbe (or B-field) with connection. In [14], Segal proposes this as the underlying geometric object of elliptic cohomology. More recently, this notion has been refined by Teichner and Stolz.

[1] Strings should also be independent of the parametrization. One should therefore take homotopy orbit spaces of the objects under the S^1 action. In that case the diffeomorphisms of the surfaces need not be the identity on the boundary. The resulting category has the same homotopy type as $\mathcal{S}X$ in the sense that its classifying space is homotopic to that of $\mathcal{S}X$.

3.2. Conformal field theory.

The category $\mathcal{S} := \mathcal{S}(*)$ is studied in conformal field theory [15]. Its objects are the natural numbers and its morphims are Riemann surfaces. A conformal field theory (CFT) is a linear space $\mathbb{H}$ which is an algebra over $\mathcal{S}$. Thus each element in $\mathcal{M}_{g,n+m}$ defines a linear map from $\mathbb{H}^{\otimes n}$ to $\mathbb{H}^{\otimes m}$. The physical states of a topological conformal field theory (TCFT) form a graded vector space A_*. Each element of the homology $H_*\mathcal{M}_{g,n+m}$ defines a linear map from $A_*^{\otimes n}$ to $A_*^{\otimes m}$.

3.3. Gromov-Witten theory.

Let X be a symplectic manifold. A model for the homotopy orbit spaces in the definition of $\mathcal{S}X$ is the fiber bundle $\mathrm{map}(F_{g,n}, X) \to \mathcal{M}_{g,n}(X) \to \mathcal{M}_{g,n}$ over the Riemann moduli space. In each fiber, F comes equipped with a complex structure and we may replace the continuous maps by the space of pseudo-holomorphic maps $\mathrm{hol}(F_g^n, X)$ yielding a category $\mathcal{S}^{\,\mathrm{hol}}X$. This is the category relevant to Gromov-Witten theory. Note, for X a complex Grassmannian, a generalized flag manifold, or a loop group, the degree d-component of $\mathrm{hol}(F_g, X)$ approximates the components of $\mathrm{map}(F_g, X)$ in homology. The categories $\mathcal{S}X$ and $\mathcal{S}^{\,\mathrm{hol}}X$ are therefore closely related.

4. From categories to infinite loop spaces

There is a functorial way to associate to a category $\mathcal{C}$ a topological space $|\mathcal{C}|$, the realization of its nerve. It takes equivalences of categories to homotopy equivalences. It is a generalization of the classifying space construction of a group: $|G| = BG$ where G is identified with the category of a single object and endomorphism set G. The path-category $\mathcal{P}X$ of a connected space X is a many object group up to homotopy. The underlying "group" is the space $\Omega X = \mathrm{map}_*(S^1, X)$ of based loops in X. Again one has $|\mathcal{P}X| \simeq B(\Omega X) \simeq X$. A functor from $\mathcal{P}X$ to n-dimensional vector spaces and their isomorphisms thus defines a map

$$X \longrightarrow BGL_n\mathbb{C},$$

and hence an element in the K-theory of X. Motivated by this, we would like to understand the classifying space of the string category $\mathcal{S}X$ and its relation to elliptic cohomology.

Definition 4.1. $St(X) := \Omega|\mathcal{S}X|$.

Theorem 4.2. *St is a homotopy functor from the category of topological spaces to the category of infinite loop spaces.*

We recall that Z is an infinite loop space if it is homotopic to some Z_0 such that successive based spaces Z_i can be found with homeomorphisms $\gamma_i : Z_i \simeq \Omega Z_{i+1}$. Any infinite loop space Z gives rise to a generalized homology theory h_* which evaluated on a space Y is

$$h_*Y := \pi_* \lim_{i\to\infty} \Omega^i(Z_i \wedge Y).$$

Infinite loop spaces are abelian groups up to homotopy in the strongest sense.

The proof of Theorem 4.2 can be sketched as follows, compare [16]. $\mathcal{S}X$ is a symmetric monoidal category under disjoint union. Infinite loop space machinery (see for example [13]) implies that its classifying space $|\mathcal{S}X|$ is a homotopy abelian monoid in the strongest sense. But $\pi_0|\mathcal{S}X| = H_1X$ is a group. Hence homotopy inverses exist and $|\mathcal{S}X|$ is an infinite loop space.

Using another piece of infinite loop space machinery (a generalization of the group completion theorem) and Harer Stability Theorem 2.1, one can identify the string theory of a point as $\mathbb{Z} \times B\Gamma_\infty^+$. As an immediate consequence we have

Corollary 4.3. *([16]) $St(*) \simeq \mathbb{Z} \times B\Gamma_\infty^+$ is an infinite loop space.*

5. CFT-operad

We offer now a different perspective on Theorem 4.1 and Corollary 4.3. Let $\mathcal{M} = \{\mathcal{M}_n\}_{n\geq 0}$ with $\mathcal{M}(n) = \coprod_{g\geq 0} B\Gamma_{g,n+1}$ be the operad contained in the CFT category $\mathcal{S}$. A space X is an algebra over $\mathcal{M}$ if there are compatible maps $\mathcal{M}(n) \times X^n \to X$. In particular X has a monoid structure. Let $\mathcal{G}X$ be its group completion. $\mathcal{G}X$ is homotopic to X if and only if $\pi_0 X$ is a group.

Theorem 5.1. *([18]) If X is an algebra over $\mathcal{M}$ then its group completion $\mathcal{G}X$ is an infinite loop space.*

CFT is therefore closely linked to the theory of infinite loop spaces. The crucial point of the proof is a decoupling result similar to Proposition 2.3. The corresponding statement for TCFT's implies that Getzler's Batalin-Vilkovisky algebra structure on the physical states A_* is stably trivial, see [17]. The following examples illustrate the strength of Theorem 5.1.

Example 5.1.

Let X_1 be the disjoint union $\coprod_{g\geq 0} B\Gamma_{g,1}$. It has a product induced by glueing $F_{g,1}$ and $F_{h,1}$ to the "legs" of a pairs of pants surface $F_{0,3}$. Miller observes in [10] that this induces a double loop space structure on the group completion $\mathcal{G}X_1 \simeq \mathbb{Z} \times B\Gamma_\infty^+$. This product extends to an $\mathcal{M}$-algebra structure. Hence Miller's double loop space structure extends to an infinite loop space structure. Wahl [19] proved that it is equivalent to the infinite loop spaces structure implied by Corollary 4.3.

Example 5.2.

Let X_2 be the disjoint union of the Borel constructions

$$E_g := E\,\mathrm{Diff}(F_{g,1}) \times_{\mathrm{Diff}(F_{g,1})} \mathrm{map}(F_{g,1}, \partial; X, *).$$

As the functions to X map the boundary to a point, they can be extended from $F_{g,1}$ to $F_{g+1,1}$ via the constant map. X_2 thus becomes an $\mathcal{M}$-algebra and $\mathcal{G}X_2 = \mathbb{Z} \times (\lim_{g\to\infty} E_g)^+$ is an infinite loop space. $\mathcal{G}X_2$ is homotopic to $\mathrm{St}(X)$ when X is simply connected.

Example 5.3.

Let $C(F_{g,1}; X)$ denote the space of unordered configurations in the interior of $F_{g,1}$ with labels in a connected space X. Let C_g be its Borel construction. Their disjoint union defines an $\mathcal{M}$-algebra X_3. The following decoupling result determines its group completion.

Proposition 5.2. $\mathcal{G}X_3 \simeq \mathbb{Z} \times (\lim_{g\to\infty} C_g)^+ \simeq \mathbb{Z} \times B\Gamma_\infty^+ \times Q(BS^1 \wedge X_+)$.

$Q = \lim \Omega^\infty S^\infty$ is the free infinite loop space functor and X_+ denotes X with a disjoint basepoint. Note the close relation with the above example. By work of McDuff and Bödigheimer, there is a homotopy equivalence

$$C(F_{g,1}; X) \simeq \text{ map}(F_{g,1}, \partial; S^2 \wedge X).$$

Note though that the induced $\text{Diff}(F_{g,1})$-action from the left on the right is non-trivial on the sphere in the target space.

6. Refinement of Mumford's conjecture

Infinite loop spaces are relatively rare and the question arises whether $\mathbb{Z} \times B\Gamma_\infty^+$ can be understood in terms of well-known infinite loop spaces. This question was addressed in joint work with Madsen.

Let $\mathbb{P}^l$ be the Grassmannian of oriented 2-planes in $\mathbb{R}^{l+2}$ and let $-L_l$ be the complement of the canonical 2-plane bundle L over $\mathbb{P}^l$. The one-point compactification, the Thom space $\text{Th}(-L_l)$, restricts on the subspace $\mathbb{P}^{l-1}$ to the suspension of $\text{Th}(-L_{l-1})$. Taking adjoints yields maps $\text{Th}(-L_{l-1}) \to \Omega\text{Th}(-L_l)$, and we may define

$$\Omega^\infty \text{Th}(-L) := \lim_{l\to\infty} \Omega^l \text{Th}(-L_l).$$

More generally, for any space X, define

$$\Omega^\infty(\text{Th}(-L) \wedge X_+) := \lim_{l\to\infty} \Omega^l(\text{Th}(-L_l) \wedge X_+).$$

Conjecture 6.1. *There is a homotopy equivalence of infinite loop space*

$$\alpha : \ St(X) \xrightarrow{\simeq} \Omega^\infty(Th(-L) \wedge X_+).$$

Remarks 6.2.

For $X = *$, Conjecture 6.1 postulates a homotopy equivalence $\alpha : \mathbb{Z} \times B\Gamma_\infty^+ \to \Omega^\infty \text{Th}(-L)$. A proof of this has been announced by Madsen and Weiss, see Section 8. The Mumford conjecture follows from this as we will explain presently. Conjecture 6.1 claims in addition that $\text{St}(\_)$ is a homology functor, i.e. $\pi_* \text{St}(\_)$ is the homology theory associated to the infinite loop space $\Omega^\infty \text{Th}(-L)$.

The infinite loop space $\Omega^\infty \mathrm{Th}(-L)$ is well-studied, more recently because of its relation to Waldhausen K-theory. The inclusion of $-L_l$ into the trivial bundle $L \oplus (-L_l) \simeq \mathbb{P}^l \times \mathbb{R}^{l+2}$ induces a map

$$\omega : \Omega^\infty \mathrm{Th}(-L) \longrightarrow Q(\mathbb{P}^\infty_+).$$

ω has homotopy fibre $\Omega^2 Q(S^0)$. As the stable homotopy groups of the sphere are torsion in positive dimensions, ω is a rational equivalence. Let $\mathbb{P}^\infty \to BU$ be the map that classifies L. By Bott periodicity, the map can be extended to the free infinite loop space $L : Q_0(\mathbb{P}^\infty) \to BU$. The subscript here indicates the 0-component. L has a splitting and is well known to be a rational equivalence:

$$H^*(\Omega^\infty_0 \mathrm{Th}(-L); \mathbb{Q}) \overset{\omega^* \circ L^*}{\simeq} \mathbb{Q}[c_1, c_2, \dots].$$

The $\mathbb{Z}/p$-homology of $\Omega^\infty \mathrm{Th}(-L)$ has recently been determined by Galatius [3].

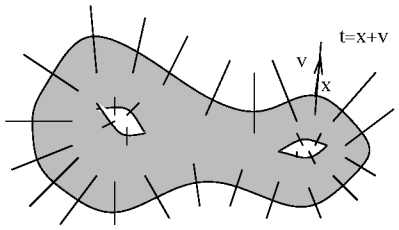


Figure 1: Surface $h(F) \subset \mathbb{R}^{l+2}$ with tubular neighborhood U.

To define a map α comes down to defining maps from the morphism spaces of $\mathcal{S}X$. (See also Example 5.2.) α is the homotopy theoretic interpretation of the formula defining κ_i where the wrong way map $\int_F$ is replaced by the (pre)transfer map of Becker and Gottlieb. We give now an explicid description of this map.

For simplicity, let F be a closed surface. Consider the space of smooth embeddings $\mathrm{Emb}(F, \mathbb{R}^{l+2})$. By Whitney's embedding theorem, as $l \to \infty$, it may serve as a model for $E\,\mathrm{Diff}(F)$. Let

$$(h, f) \in \mathrm{Emb}(F, \mathbb{R}^{l+2}) \times_{\mathrm{Diff}(F)} \mathrm{map}(F, X).$$

Choose a tubular neighborhood U of $h(F)$ such that every $t \in U$ can uniquely be written as $x + v$ with $x \in h(F)$ and v normal to the tangent plane $T_x h(F)$. α sends (h, f) to the continuous function $\alpha(h, f) : S^{l+2} \to \mathrm{Th}(-L_l) \wedge X_+$ defined by

$$t \quad \mapsto \quad \begin{cases} \infty & \text{if } t \notin U, \\ ((T_x h(F), v), f(h^{-1}(x)) & \text{if } t = x + v \in U. \end{cases}$$

In [8], α is shown to be a 3-connected map of infinite loop spaces and the tautological classes are identified. Let $i!(ch_i) \in H^{2i} BU$ denote the i-th integral Chern character class. Then

$$\kappa_i = \alpha^* \circ \omega^* \circ L^*(i! ch_i).$$

7. Splittings and (co)homological results

The main result of [8] is a partial splitting of the composition

$$\omega \circ \alpha : \mathbb{Z} \times B\Gamma_\infty^+ \longrightarrow Q(\mathbb{P}_+^\infty) \simeq Q(S^0) \times Q(\mathbb{P}^\infty).$$

This is achieved by constructing a map μ from $\mathbb{P}_+$ to $\mathbb{Z} \times B\Gamma_\infty^+$ and then extending it to the free infinite loop space $Q(\mathbb{P}_+^\infty)$ utilizing the infinite loop space structure on $\mathbb{Z} \times B\Gamma_\infty^+$. In order to construct μ, approximate $\mathbb{P}^\infty \simeq BS^1$ by the classifying spaces of cyclic groups C_{p^n} for $n \to \infty$, one prime p at a time, as the cyclic groups can be mapped into suitable mapping class groups. However, this means that we have to work with p-completions.

Let $Y_p^\wedge$ denote the p-completion of Y and $g \in \mathbb{Z}_p^\times$ be a topological generator of the p-adic units ($g = 3$ if $p = 2$). Denote by $\psi^k : \mathbb{P}^\infty \to (\mathbb{P}^\infty)_p^\wedge$ the map that represents k times the first Chern class in $H^2(\mathbb{P}^\infty, \mathbb{Z}_p)$.

Theorem 7.1. **[8].** *There exists a map* $\mu : (Q(S^0) \times Q(\mathbb{P}^\infty))_p^\wedge \to (\mathbb{Z} \times B\Gamma_\infty^+)_p^\wedge$ *such that*

$$\omega \circ \alpha \circ \mu \simeq \begin{pmatrix} -2 & * \\ 0 & 1 - g\psi^g \end{pmatrix}.$$

The map $1 - g\psi^g$ induces multiplication by $1 - g^{n+1}$ on $H_{2n}(\mathbb{P}^\infty; \mathbb{Z}_p)$ which is a p-adic unit precisely if $n \neq -1(\mod p-1)$. The following applications of Theorem 7.1 are also found in [8]. There is a splitting $Q(\mathbb{P}_+^\infty)_p^\wedge \simeq E_0 \times \cdots \times E_{p-2}$ corresponding to the idempotent decomposition of $\mathbb{Z}[\mathbb{Z}/p^\times] \subset \mathbb{Z}_p[\mathbb{Z}_p^\times]$.

Corollary 7.2. *For some* W_p, *there is a splitting of infinite loop spaces*

$$(\mathbb{Z} \times B\Gamma_\infty^+)_p^\wedge \simeq E_0 \times \cdots \times E_{p-3} \times W_p.$$

The $\mathbb{Z}/p$-homology of $Q(\mathbb{P}_+^\infty)$ is well-understood in terms of Dyer-Lashof operation. These are homology operations for infinite loop spaces that are formally similar to the Steenrod operations. For each generator $a_i \in H_{2i}\mathbb{P}^\infty = \mathbb{Z}$ there is an infinite family of $\mathbb{Z}/p$-homology classes freely generated by the Dyer-Lashof operations. The product $E_0 \times \cdots \times E_{p-3}$ contains precisely those families for which $i \neq -1(\mathrm{mod}\, p-1)$, giving a huge collection of new p-torsion in $H_*B\Gamma_\infty$.

For odd primes p, Madsen and Schlichtkrull [MS] found split surjective maps l_0 and l_{-1} of infinite loop spaces such that the following diagram is commutative

$$\begin{array}{ccc} \Omega^\infty \mathrm{Th}(-L)_p^\wedge & \xrightarrow{\omega} & Q(\mathbb{P}_+^\infty)_p^\wedge \\ {\scriptstyle l_{-1}}\downarrow & & {\scriptstyle l_0}\downarrow \\ (\mathbb{Z} \times BU)_p^\wedge & \xrightarrow{1-g\psi^g} & (\mathbb{Z} \times BU)_p^\wedge. \end{array}$$

Corollary 7.3. *For odd primes p and some space V_p, there is a splitting of spaces*

$$(B\Gamma_\infty^+)_p^\wedge \simeq BU_p^\wedge \times V_p.$$

This gives a $\mathbb{Z}_p$-integral version of Miller and Morita's theorem: the polynomial algebra $\mathbb{Z}_p[c_1, c_2, \dots]$ is a split summand of $H^*(B\Gamma_\infty; \mathbb{Z}_p)$. The divisibility of the tautological classes κ_i at odd primes p can also be deduced from the above diagram.

Corollary 7.4. *If $i = -1 (mod\ p-1)$, then κ_i is divisible by $p^{1+\nu_p(i+1)}$ where ν_p is the p-adic valuation. Otherwise, p does not divide κ_i.*

In the light of [9] this result is sharp.

8. Geometric interpretation

$\alpha : B\Gamma_\infty^+ \to \Omega_0^\infty \mathrm{Th}(-L)$ is a homotopy equivalence if and only if it induces an isomorphism in oriented cobordism theory Ω_*^{SO}. An element in $\Omega_n^{\mathrm{SO}}(B\Gamma_\infty^+) = \Omega_n^{\mathrm{SO}}(B\Gamma_\infty)$ is a cobordism class of oriented surface bundles $F \to E^{n+2} \xrightarrow{\pi} M^n$. An element in $\Omega_n^{\mathrm{SO}}(\Omega_0^\infty \mathrm{Th}(-L))$ is a cobordism class of pairs $[\pi : E^{n+2} \to M^n, \hat{\pi}]$ of smooth maps π and stable bundle surjections from TE to π^*TM. (Upto cobordism one can assume that $\hat{\pi}$ is a vector bundle surjection.) α maps a bundle $[F \to E \xrightarrow{\pi} M]$ to the pair $[\pi : E \to M, D\pi]$ where $D\pi$ denotes the differential of π. Hence, α is a homotopy equivalence if and only if each cobordism class of pairs $[\pi : E^{n+2} \to M^n, \hat{\pi}]$ contains a "unique" representative with π a submersion.

It is this geometric formulation that underpins the solution to the Mumford conjecture by Madsen and Weiss. A key ingredient of the proof is the Phillips-Gromov h-principle of submersion theory: A pair $(g : X \to M, \hat{g} : TX \to g^*TM)$ can be deformed to a submersion – provided X is open. E above, however, is closed. The approach taken in [9] is to replace $\pi : E \to M$ by $g = \pi \circ pr_1 : X = E \times \mathbb{R} \to M$. Now the submersion h-principle applies and g can be replaced by a submersion f. The proof then consists of a careful analysis of the singularities of the projection $pr_1 : X \to \mathbb{R}$ on the fibers of f. At a critical point it uses Harer's Stability Theorem 2.1.

Madsen-Weiss Theorem 8.1. *The map $\alpha : \mathbb{Z} \times B\Gamma_\infty^+ \to \Omega^\infty Th(-L)$ is a homotopy equivalence.*

References

[1] C.-F. Bödigheimer & U. Tillmann, *Stripping and splitting decorated mapping class groups*, Birkhäuser, Progress in Math. **196** (2001), 47– 57.

[2] C.J. Earle & J. Eells, *A fibre bundle description of Teichmüller theory*, J. Diff. Geom. **3** (1969), 19–43.

[3] S. Galatius, *Homology of $\Omega^\infty \Sigma \mathbb{C}P_{-1}^\infty$ and $\Omega^\infty \mathbb{C}P_{-1}^\infty$*, preprint 2002.

[4] J.L. Harer, *Stability of the homology of the mapping class groups of orientable surfaces*, Annals Math. **121** (1985), 215–249.

[5] N.V. Ivanov, *Stabilization of the homology of Teichmüller modular groups*, Leningrad Math. J. **1** (1990), 675–691.

[6] N.V. Ivanov, *On the homology stability for Teichmüller mudular groups: closed surfaces and twisted coefficients*, Mapping Class Groups and Moduli Spaces of Riemann Surfaces, Contemp. Math. **150** (1993), 149–194.

[7] I. Madsen & C. Schlichtkrull, *The circle transfer and K-theory*, AMS Contemporary Math. **258** (2000), 307–328.

[8] I. Madsen & U. Tillmann, *The stable mapping class group and* $Q(\mathbb{C}P^\infty)$, Invent. Math. **145** (2001), 509–544.

[9] I. Madsen & M. Weiss, *Cohomology of the stable mapping class group*, in preparation.

[10] E.Y. Miller, *The homology of the mapping class group*, J. Diff. Geom. **24** (1986), 1–14.

[11] S. Morita, *Characteristic classes of surface bundles*, Invent. Math. **90** (1987), 551-577.

[12] D. Mumford, *Towards an enumerative geometry of the moduli space of curves*, Arithmetic and Geometry, M. Artin and J. Tate, editors, Progr. Math., Birkhauser **36** (1983), 271–328.

[13] G. Segal, *Categories and cohomology theories*, Topology **13** (1974), 293–312.

[14] G. Segal, *Elliptic cohomology (after Landweber-Stong, Ochanine, Witten, and others)*, Seminar Bourbaki, Asterisque **161-162** (1989), 187–201.

[15] G. Segal, *The definition of conformal field theory*, manuscript.

[16] U. Tillmann, *On the homotopy of the stable mapping class group*, Invent. Math. **130** (1997), 257–275.

[17] U. Tillmann, *Vanishing of the Batalin-Vilkovisky algebra structure for TCFTs*, Commun. Math. Phys **205** (1999), 283–286.

[18] U. Tillmann, *Higher genus surface operad detects infinite loop spaces*, Math. Ann. **317** (2000), 613–628.

[19] N. Wahl, *Infinite loop space structure(s) on the stable mapping class group*, Oxford Thesis 2001.

ICM 2002 · Vol. II · 457–468

Non-zero Degree Maps between 3-Manifolds*

Shicheng Wang†

Abstract

First the title could be also understood as "3-manifolds related by non-zero degree maps" or "Degrees of maps between 3-manifolds" for some aspects in this survey talk.

The topology of surfaces was completely understood at the end of 19-th century, but maps between surfaces kept to be an active topic in the 20-th century and many important results just appeared in the last 25 years. The topology of 3-manifolds was well-understood only in the later 20-th century, and the topic of non-zero degree maps between 3-manifolds becomes active only rather recently.

We will survey questions and results in the topic indicated by the title, present its relations to 3-manifold topology and its applications to problems in geometry group theory, fixed point theory and dynamics.

There are four aspects addressed: (1) Results concerning the existence and finiteness about the maps of non-zero degree (in particular of degree one) between 3-manifolds and their suitable correspondence about epimorphisms on knot groups and 3-manifold groups. (2) A measurement of the topological complexity on 3-manifolds and knots given by "degree one map partial order", and the interactions between the studies of non-zero degree map among 3-manifolds and of topology of 3-manifolds. (3) The standard forms of non-zero degree maps and automorphisms on 3-manifolds and applications to minimizing the fixed points in the isotopy class. (4) The uniqueness of the covering degrees between 3-manifolds and the uniqueness embedding indices (in particular the co-Hopfian property) between Kleinian groups .

The methods used are varied, and we try to describe them briefly.

2000 Mathematics Subject Classification: 57M, 55C, 37E, 30F40, 20E26.

0. Introduction

The topology of surfaces was completely understood by the end of 19th century, but maps between surfaces keep to be an active topic in the 20th century, and

*Supported by grants of MSTC and NSFC.

†Department of Mathematics, Peking University, Beijing 100871, China. E-mail: swang@sxx0.math.pku.edu.cn

some basic results just appeared in the last 25 years, among which are the Nielsen-Thurston classification of surface automorphisms [Th3], and Edmonds' standard form for surface maps [E1]. Then fine results followed, say the realization of Nielsen number in the isotopy class of surface automorphisms by Jiang [J], and the simple loop theorem for surface maps by Gabai [Ga2].

The topology of 3-manifolds was well-understood only in the later 20th century due to many people's deep results, in particular Thurston's great contribution to the geometrization of 3-manifolds, and the topic of non-zero degree maps between 3-manifolds becomes active only rather recently.

We will survey the results and questions in the topic indicated by the title, present its relations to 3-manifold topology and its applications to problems in geometry group theory, fixed point theory and dynamics. The methods used are varied, and we try to describe them briefly.

For standard terminologies of 3-manifolds and knots, see the famous books of J. Hempel, W. Jaco and D. Rolfsen. For a proper map $f : M \to N$ between oriented compact 3-manifolds, $\deg(f)$, the degree of f, is defined in most books of algebraic topology. A closed orientable 3-manifold is said to be *geometric* if it admits one of the following geometries: H^3 (hyperbolic), $\widetilde{PSL_2R}$, $H^2 \times E^1$, Sol, Nil, E^3 (Euclidean), $S^2 \times E^1$, and S^3 (spherical). A compact orientable 3-manifold M admits a *geometric decomposition* if each prime factor of M is either geometric or Haken. Thurston's geometrization conjecture asserts that any closed orientable 3-manifold admits a geometric decomposition. Each Haken manifold M with ∂M a (possibly empty) union of tori has a Jaco-Shalen-Johannson (JSJ) torus decomposition, that is, it contains a minimal set of tori, unique up to isotopy, cutting M into pieces such that each piece is either a Seifert manifold or a simple manifold, which admits a complete hyperbolic structure with finite volume [Th2].

In the remainder of the paper, all manifolds are assumed to be **compact and orientable**, all automorphisms are **orientation preserving**, all knots are **in** S^3, all Kleinian groups are **classical**, and all maps are **proper,** unless otherwise specified.

Let M and N be 3-manifolds and $d > 0$ an integer. We say that M *d-dominates* (or simply dominates) N if there is a map $f : M \to N$ of degree $\pm d$. Denote by $D(M, N)$ the set of all possible degrees of maps from M to N. A 3-manifold M is *small* if each closed incompressible surface in M is boundary parallel.

Due to space limitation, quoted literature are only partly listed in the references; while the others are briefly indicated in the context.

1. Existence and finiteness

A fundamental question in this area (and in 3-manifold theory) is the following.

Question 1.1. Given a pair of closed 3-manifolds M and N, can one decide if M d-dominates N? In particular, can one decide if M 1-dominates N?

The following two natural problems concerning finiteness can be considered as testing cases of Question 1.1.

Question 1.2 [Ki, Problem 3.100 (Y. Rong)]. Let M be a closed 3-manifold. Does M 1-dominate at most finitely many closed 3-manifolds?

Question 1.3. Let N be a closed 3-manifold. When is $|D(M,N)|$ finite for any closed 3-manifold M?

An important progress towards the solution of Question 1.2 is the following:

Theorem 1.1 ([So2], [WZh2], [HWZ3]). *Any closed 3-manifold 1-dominates at most finitely many geometric 3-manifolds.*

Theorem 1.1 was proved by Soma when the target manifolds N admit hyperbolic geometry [So2]. The proof is based on the argument of Thurston's original approach to the deformation of acylindrical manifolds. Porti and Reznikov had a quick proof of Soma's result, based on the volume of representations [Re2]. However Soma's approach deserves attention as it proves that the topological types of all hyperbolic pieces in closed Haken manifolds 1-dominated by M are finite [So3]. Theorem 1.1 was proved in [WZh2] when the target manifolds admit geometries of $H^2 \times E^1$, $\widetilde{PSL_2(R)}$, Sol or Nil. The proof for the case of $H^2 \times E^1$ geometry invokes Gabai's result that embedded Thurston Norm and singular Thurston Norm are equal [Ga1], and the proof for case of $\widetilde{PSL_2(R)}$ geometry uses Brooks and Goldman's work on Seifert Volume [BG]. Theorem 1.1 was proved in [HWZ3] when the target manifolds admit S^3 geometry, using the linking pairing of 3-manifolds. Note that only finitely many 3-manifolds admit the remaining two geometries.

For maps between 3-manifolds which are not necessarily orientable, there is a notion of geometric degree (See D. Epstein, Proc. London Math. Soc. 1969). It is worth mentioning that if d-dominating maps are defined in terms of geometric degree, then Rong constructs a non-orientable 3-manifold which 1-dominates infinitely many lens spaces [Ro3]. Actually there is a non-orientable hyperbolic 3-manifold which 1-dominates infinitely many hyperbolic 3-manifolds [BW1]. Such examples do not exist in dimension $n > 3$ due to Gromov's work on simplicial volume and H.C. Wang's theorem that, for any $V > 0$, there are at most finitely many closed hyperbolic n-manifolds of volume $< V$.

The answer to Question 1.2 is still unknown for closed irreducible 3-manifolds admitting geometric decomposition. The following result is related.

Theorem 1.2 ([Ro1], [So3]). *For any 3-manifold M there exists an integer N_M, such that if $M = M_0 \to M_1 \to .... \to M_k$ is a sequence of degree one maps with $k > N_M$, and each M_i admits a geometric decomposition, then the sequence contains a homotopy equivalence.*

The situation for Question 1.3 can be summarized in the following theorem.

Theorem 1.3 ([Gr1], [BG], [W2]). *Suppose N is a closed 3-manifold admitting geometric decomposition. Then*

(1) $|D(M,N)|$ is finite if either a prime factor of N contains a hyperbolic piece in its JSJ decomposition, or N itself admits the geometry of $\widetilde{PSL_2(R)}$.

(2) $|D(N,N)|$ is infinite if and only if either (i) N is covered by a torus bundle over the circle or a surface$\times S^1$, or (ii) each prime factor of N has a cyclic or finite fundamental group.

Part (1) of Theorem 1.3 follows from the work of Gromov [Gr1] and Brooks-Goldman [BW]. Part (2) can be found in [W2]. Note that if $|D(N,N)|$ is infinite and $D(M,N)$ contains non-zero integers, then $|D(M,N)|$ is also infinite. I suspect that Theorem 1.3 (2) indicates a general solution to Question 1.3.

There are many partial results for Question 1.1: When both M and N are Seifert manifolds with infinite fundamental groups Rong has an algorithm to determine if M 1-dominates N [Ro3]. When N is the Poincare homology sphere and a Heegaard diagram of M is given, Hayat-Legrand, Matveev and Zieschang have an algorithm to decide if M d-dominates N [HMZ]. There are simple answers to Question 1.1 in the following cases: (1) M and N are prism spaces and $d = 1$ [HWZ2]; (2) $M = N$ admit geometry of S^3 and f_* an automorphism on π_1 [HKWZ]; (3) N is a lens space. I will state (3) as a theorem, since both its statement and proof are short, and since it has rich connections with previous results and with different topics.

Theorem 1.4 ([HWZ1], [HWZ3]). *A closed 3-manifold M d-dominates the lens space $L(p,q)$ if and only if there is an element α in the torsion part of $H_1(M,\mathbb{Z})$ such that $\alpha \odot \alpha = \frac{dq}{p}$ in $\mathbb{Q}/\mathbb{Z}$, where $\alpha \odot \alpha$ is the self-linking number of α.*

A direct consequence of Theorem 1.4 is the known fact that $L(p,q)$ 1-dominates $L(m,n)$ if and only if $p = km$ and $n = kqc^2 \bmod m$. This fact has at least four different proofs: using equivariant maps between spheres by de Rham (J. Math. 1931) and by Olum (Ann. of Math. 1953), using Whitehead torsion by Cohen (GTM 10, 1972), using pinch in [RoW] and using linking pair in [HWZ1].

Degree one maps from general 3-manifolds to some lens spaces, in particular the RP^3, have been studied by Bredon-Wood (Invent. Math. 1969) and by Rubinstein (Pacific J. Math. 1976) to find one-sided incompressible surfaces, by Luft-Sjerve (Topo. Appl. 1990) to study cyclic group actions on 3-manifolds, by Shastri-Williams-Zvengrowski [SWZ] in theoretical physics, by Taylor (Topo. Appl. 1984) to define normal bordism classes of degree one maps, and by Kirby-Melvin (Invent. Math. 1991) to connect with new 3-manifold invariants.

Degree one maps induce epimorphisms on π_1. There are easy examples indicate that Question 1.2 does not have direct correspondence in the level of 3-manifold groups [BW1], [RWZh]. However the following related question was raised in 1970's.

Question 1.4 [Ki, Problem 1.12 (J. Simon)]. Conjectures:

(1) Given a knot group G, there is a number N_G such that any sequence of epimorphisms of knot groups $G \to G_1 \to .... \to G_n$ with $n \geq N_G$ contains an isomorphism.

(2) Given a knot group G, there are only finitely many knot groups H for which there is an epimorphism $G \to H$.

According to a conversation with Gonzalez-Acuna, who discussed Question 1.4 with Simon before it was posed, the epimorphisms in Question 1.4 are peripheral preserving in their minds.

Theorem 1.5 ([So5], [RW]). *The conjecture in Question 1.4 (1) holds if the knot complements involved are small. The conjecture in Question 1.4 (2) holds if the knot complements are small and the epimorphisms are peripheral preserving.*

The first claim is due to Soma [So5] and the second claim is in [RW]. Both of them invoke Culler-Shalen's work on the representation varieties of knot groups. It is also proved that any infinite sequence of epimorphisms among 3-manifold groups contains an isomorphism if all manifolds are either hyperbolic [So5] or Seifert fibered [RWZh]. In [RWZh], the proof uses the fact that epimorphisms between aspherical Seifert manifolds with the same π_1 rank are realized by maps of non-zero degree. Both this fact and Question 1.4 (1) are variations of the Hopfian property.

We end this section by mention that there are results about $D(M, N)$ in [DW] for $(n-1)$-connected $2n$-manifolds, $n > 1$, which are quite explicit and of interest from both topological and number-theoretic point of view.

2. Uniqueness

The following question is raised in 1970's.

Question 2.1 [Ki, Problem 3.16 (W. Thurston)]. Suppose a 3-manifold M is not covered by (surface)$\times S^1$ or a torus bundle over S^1. Let $f, g : M \to N$ be two coverings, must $deg(f) = deg(g)$?

It is known [WWu2] that Question 2.1 has positive answer if M admits geometric decomposition and is not a graph manifold (M is a graph manifold if each piece of its JSJ decomposition is Seifert fibered.) For graph manifolds there are four different covering invariants introduced in middle 1990's by [WWu2], Luecke and Wu [LWu], Neumann [N] and Reznikov [Re1] . Unfortunately all those four covering invariants are either vanishing or not well-defined for some non-trivial graph manifolds. It is also known that covering degree is uniquely determined if the graph manifold in the target is either a knot complement [LWu] or its corresponding graph is simple [WWu2, N]. The positive answer to Question 2.1 for graph manifolds was finally obtained in [YW], using the matrix invariant defined in [WWu2] and an elegant application of matrix theory due to Yu.

Theorem 2.1 ([WWu2], [YW]). *For 3-manifolds admitting geometric decomposition and not covered by either (surface)$\times S^1$ or a torus bundle over S^1, covering degrees are uniquely determined by the manifolds involved.*

It is worth mentioning an interesting fact that any knot complement is non-trivially covered by at most two knot complements and any knot complement non-trivially covers at most one knot complement. The first claim follows from the cyclic surgery theorem of Culler-Gordon-Luecke-Shalen and the positive answer to the Smith Conjecture. The second claim is in [WWu1].

Question 2.1 is equivalent to asking the uniqueness of indices of finite index embeddings between 3-manifold groups. Recently there are also some discussions on the uniqueness of indices of self-embeddings of groups. A group G is said to be co-Hopf if each self-monomorphism of G is an isomorphism.

Question 2.2. Let G be either a 3-manifold group, or a Kleinian group, or a word hyperbolic group. When is G co-Hopf?

The cohopficity of groups were first considered by Baer (Bull. AMS 1944). For word hyperbolic groups it was first considered by Gromov in 1987 [Gr2, p.157], and subsequently by Rips-Sela (GAFA, 1994), Sela [Se], and Kapovich-Wise (Isreal J. Math. 2001). Cohopficity of 3-manifold groups was first studied in 1989 by Gonzalez-Acuna and Whitten [GWh], and then in [WWu2] and [PW]. The answer for 3-manifolds admitting geometric decomposition with boundary either empty set or a union of tori is known [GWh], [WWu2], and partial results for 3-manifolds with boundary of high genus surfaces are in [PW]. Cohopficity of Kleinian groups was first considered in 1992 in an early version of [PW], then in 1994 in an early version of [WZh1], also by Ohshika-Potyagailo (Ann. Sci. Ecole Norm. Sup. 1998) and Delzant-Potyagailo (MPI Preprint, 2000) for high dimensional Kleinian groups.

Theorem 2.2. *Suppose K is a non-elementary, freely indecomposable, geometrically finite Kleinian group and K contains no $Z \oplus Z$ subgroup. Then*

(1)[Se], [PW], [WZh1] *K is co-Hopf if K is a group of one end.*

(2)[WZh1] *If the singular locus of the hyperbolic 3-orbifold H^3/K is a 1-manifold, then K is co-Hopf if and only if no circle component of singular locus meets a minimal splitting system of hyperbolic cone planes.*

The proof of Theorem 2.2 (1) in [WZh1], influenced by that of torsion free case in [PW], use a generalization of Thurston-Gromov's finiteness theorem on the conjugacy classes of group embeddings (Delzant, Comm. Math. Helv. 1995) and a proper conjugation theorem of Kleinian groups (Wang-Zhou, Geometriae Dedicata, 1995). Theorem 2.2 (2) is proved by using 3-dimensional hyperbolic orbifold structures and orbifold maps, which turn out to be useful geometric tools.

Note that groups in Theorem 2.2 are word hyperbolic groups. According to Sela ([Se] and his MSRI preprints in 1994), people once expected that a non-elementary word hyperbolic group is co-Hopf if and only if it has one end. Sela proved this expectation for the torsion free case [Se]. Theorem 2.2 (2) and examples

in [WZh1] show that cohopficity phenomenon is very complicated in the torsion case. In particular there are co-Hopf word hyperbolic groups which have infinitely many ends.

Inspired by Questions 2.1 and 2.2 it is natural to ask

Question 2.3. Are the indices (including the infinity) of embeddings $H \to G$ between co-Hopf groups unique?

3. Interactions with 3-manifold topology

Degree one maps define a partial order on Haken manifolds and hyperbolic 3-manifolds. By Gordon-Luecke's theorem knots are determined by their complements [GL]. We say that a knot K 1-dominates a knot K' if the complement of K 1-dominates the complement of K'. 1-domination among knots also gives a partial order on knots. This partial order seems to provide a good measurement of complexity of 3-manifolds and knots. The reactions of non-zero degree maps between 3-manifolds and 3-manifold topology are reflected in the following very flexible

Question 3.1. Suppose M and N are 3-manifolds (knots) and M 1-dominates (d-dominates) N.

(1) Is $\sigma(M)$ not "smaller" than $\sigma(N)$ for a topological invariant σ?

(2) If M and N are quite "close", are they homeomorphic? do they admit the same topological structure?

Positive answers to Question 3.1 (1) are known in many cases. Suppose M 1-dominates N. Then $\sigma(M) \geq \sigma(N)$ when σ is either the rank of π_1, or Gromov's simplicial volume, or Haken number (of incompressible surfaces), or genus of knots; $\sigma(N)$ is a direct summand of $\sigma(M)$ when σ is the homology group, and $\sigma(N)$ is a factor of $\sigma(M)$ if σ is the Alexander polynomial of knots. The answer to Question 3.1 (1) is still unknown for many invariants of knots and 3-manifolds, for example crossing number, unknotting number, Jones polynomial, knot energy, and tunnel number, etc. Li and Rubinstein are specially interested in Question 3.1 (1) for Casson invariant in order to prove it is a homotopy invariant [LRu].

There are both positive and negative answers to Question 3.1 (2), depending on the interpretation of the problem. On the negative side, Kawauchi has constructed, using the imitation method invented by himself, degree one maps between non-homeomorphic 3-manifolds M and N with many topological invariants identical, see his survey paper [Ka]. On the positive side, there are many results. An easy one is that if M d-dominates N and both M and N are aspherical Seifert manifolds, then the Euler number of M is zero if and only if that of N is zero [W1]. A deeper result is Gromov-Thurston's Rigidity theorem, which says that a degree one map between hyperbolic 3-manifolds of the same volume is homotopic to an isometry [Th2]. The following are some recent results in this direction.

Theorem 3.1 ([So4], [So1]). *(1) For any $V > 0$, suppose $f : M \to N$ is a degree one map between closed hyperbolic 3-manifolds with $Vol(M) < V$. Then there is a*

constant $c = c(V)$ such that $(1-c)Vol(M) \leq Vol(N)$ implies that f is homotopic to an isometry.

(2) If $M \to N$ is a map of degree d between Haken manifolds such that $||M|| = d||N||$, then f can be homotoped to send $H(M)$ to $H(N)$ by a covering, where $||*||$ is the Gromov norm and $H(*)$ is the hyperbolic part under the JSJ decomposition.

Theorem 3.2 ([BW1], [BW2]). *(1) Let M and N be two closed irreducible 3-manifolds with the same first Betti number and suppose M is a surface bundle. If $f : M \to N$ is a map of degree d, then N is also a surface bundle.*

(2) Let M and N be two closed, small hyperbolic 3-manifolds. If there is a degree one map $f : M \to N$ which is a homeomorphism outside a submanifold $H \subset N$ of genus smaller than that of N, then M and N are homeomorphic.

(1) and (2) of Theorem 3.1 provide a stronger version and a generalization of Gromov-Thurston's Rigidity theorem, respectively. In respect of Theorems 3.2, the following examples should be mentioned: There are degree one maps between two non-homeomorphic hyperbolic surface bundles with the same first Betti number and between two non-homeomorphic small hyperbolic 3-manifolds [BW2]. The constructions of those maps are quite non-trivial. There are many applications of Theorem 3.2. We list two of them which are applications of Theorem 3.2 to Thurston's surface bundle conjecture and to Dehn surgery respectively, where degree one maps constructed by surgery on null-homotopic knots are involved.

Theorem 3.3 ([BW1], [BW2]). *(1) There are closed hyperbolic 3-manifolds M such that any tower of abelian covering of M contains no surface bundle.*

(2) Suppose M is a small hyperbolic 3-manifold and that $k \subset M$ is a null-homotopic knot, which is not in a 3-ball. If the unknotting number of k is smaller than the Heegaard genus of M, then every closed 3-manifold obtained by a non-trivial Dehn surgery on k contains an incompressible surface.

4. Standard forms

Question 4.1. What are standard forms of non-zero degree maps and of automorphisms of 3-manifolds?

Sample answers to analogs of Question 4.1 in dimension 2 are that each map of non-zero degree between closed surfaces is homotopic to a pinch followed by a branched covering [E1], and each automorphism on surfaces can be isotoped to a map which is either pseudo Anosov (Anosov), or periodic, or reducible [Th3].

Theorem 4.1 (Haken, Waldhausen, [E2], [Ro2]). *(1) A degree one map between closed 3-manifolds is homotopic to a pinch.*

(2) A map of degree at least three between closed 3-manifolds is homotopic to a branched covering.

(3) A non-zero degree map between Seifert manifolds with infinite π_1 is homotopic to a fiber preserving pinch followed by a fiber preserving branched covering.

(1) is proved by Haken (Illinois J. Math. 1966), also by Waldhausen, and a quick proof using differential topology is in [RoW]. (2) is proved by Edmonds [E2] quickly after Hilden-Montesinos's result that each 3-manifold is a 3-fold branched covering of 3-sphere. (3) is due to Rong [Ro2], which invokes [E1]. According to conversations with D. Gabai and with M. Freedman, people are still wondering if each map of degree 2 between closed 3-manifolds is homotopic to a pinch followed by a double branched covering.

For non-prime 3-manifolds, Cesar de Sa and Rourke claim that every automorphism is a composition of those preserving and permuting prime factors (Bull. AMS, 1979), and those so-called sliding maps. A proof is given by Hendricks and Laudenbach [HL], and by McCullough [Mc].

Standard forms of automorphisms on prime 3-manifolds admitting geometric decomposition have been studied in [JWW]. The orbifold version of Nielsen-Thurston's classification of surface automorphisms is established, i.e., each orbifold automorphism is orbifold-isotopic to a map which is either (pseudo) Anosov, or periodic, or reducible. We then have the following theorem.

Theorem 4.2 ([JWW]). *Let M be a closed prime 3-manifold admitting geometric decomposition. Let $f : M \to M$ be an automorphism. Let $\mathcal{T}$ be the product neighborhood of the JSJ tori. Then*

(1) f is isotopic to an affine map if M is a 3-torus.

(2) f is isotopic to an isometry if M is the Euclidean manifold having a Seifert fibration over RP^2 with two singular points of index 2.

(3) f is isotopic to a map which preserves the torus bundle structure over 1-orbifold if M admits the geometry of Sol.

(4) for all the remaining cases, f can be isotoped so that $\mathcal{T}$ is invariant under f, and for each f-orbit O of the components in $\{\mathcal{T}, \overline{M-\mathcal{T}}\}$, $f|O$ is an isometry if O is hyperbolic, $f|O$ is affine if O belongs to $\mathcal{T}$, otherwise there is a Seifert fibration on O so that f is fiber preserving and the induced map on the orbifold is either periodic, or (pseudo) Anosov, or reducible.

As in dimension 2, standard forms in Theorems 4.2 are useful in the study of fixed point theory and dynamics of 3-manifold automorphisms. The following is a result in this direction.

Theorem 4.3 ([JWW]). *Suppose M is a closed prime 3-manifold admitting geometric decomposition and $f : M \to M$ is an automorphism. Then*

(1) the Nielsen number $N(f)$ is realized in the isotopy class of f.

(2) f is isotopic to a fixed point free automorphism unless some component of the JSJ decomposition of M is a Seifert manifold whose orbifold is neither a 2-sphere with a total of at most three holes or cone points nor a projective plane with a total of at most two holes or cone points.

References

[BW1] M. Boileau and S.C. Wang, *Non-Zero degree maps and surface bundles over*

S^1, J. Diff. Geom. **43** (1996), 789–908.

[BW2] M. Boileau and S.C.Wang, *Degree one maps, incompressible surfaces and Heegaard genus*, Preprint (2002).

[BG] R. Brooks and W. Goldman, *Volume in Seifert space*, Duke Math. J. **51** (1984), 529–545.

[CS] M. Culler and P. Shalen, *Varieties of group representations and splittings of 3-manifolds*, Ann. of Math. **117** (1983), 109–146.

[DW] H. Duan and S.C.Wang, *The degree of maps between manifolds*, Math. Zeit. (to appear).

[E1] A. Edmonds, *Deformation of maps to branched covering in dimension 2*, Ann. of Math. **110** (1979), 113–125.

[E2] ———, *Deformation of maps to branched covering in dimension 3*, Math. Ann. **245** (1979), 273–279.

[Ga1] D. Gabai, *Foliations and the topology of 3-manifolds*, J. Diff. Geom. **18** (1983), 479–536.

[Ga2] ———, *Simple loop theorem*, J. Diff. Geom. **21** (1985), 143–149.

[GL] C. Gordon and J. Luecke, *Knots are determined by their complements*, JAMS **2** (1989), 371–415.

[GWh] F. Gonzalez-Acuna and W. Whitten, *Embeddings of 3-manifold groups*, Mem. AMS **474**.

[Gr1] M. Gromov, *Volume and bounded cohomology*, Publ. Math. IHES **56** (1983), 5–99.

[Gr2] ———, *Hyperbolic groups*, Essays in Group Theory, edited by S.M. Gersten, MSRI Pub., vol. 8, Springer-Verlag, 75–263.

[HMZ] C. Hayat-Legrand, S. Matveev and H. Zieschang, *Computer calculation of the degree of maps into the Poincare homology sphere*, Experiment. Math. **10** (2001), 497–508.

[HKWZ] C. Hayat-Legrand, E. Kudryavtseva, S.C.Wang and H. Zieschang, *Degrees of self-mappings of Seifert manifolds with finite* π_1, Rend. Istit. Mat. Univ. Trieste **32** (2001), 131–147.

[HWZ1] C. Hayat-Legrand, S.C.Wang and H. Zieschang, *Degree one map onto lens spaces*, Pacific J. Math. **176** (1996), 19–32.

[HWZ2] ———, *Minimal Seifert manifolds*, Math. Ann. **308** (1997), 673–700.

[HWZ3] ———, *Any 3-manifold 1-dominates only finitely many 3-manifolds supporting* S^3 *geometry*, Proc. AMS **130** (2002), 3117–3123.

[HL] H. Hendriks and F. Laudenbach, *Diffeomorphismes des sommes connexes en dimension trois.*, Topology **23** (1984), 423–443.

[J] B. Jiang, *Fixed points of surface homeomorphisms*, Bull. AMS **5** (1981), 176–178.

[JWW] B. Jiang, S.C.Wang, and Y-Q.Wu, *Homeomorphisms of 3-manifolds and the realization of Nielsen number*, Comm. Anal. Geom. **9** (2001), 825–877.

[Ka] A. Kawauchi, *Topological imitations*, Lectures at Knots 96 (ed. Shin'ichi Suzuki), World Sci. Publ. Co., 1997, 19–37.

[Ki] R. Kirby, *Problems in low-dimensional topology,*, Geometric topology, (ed. H.Kazez), AMS/, International Press, 1997, 35–473.

[LRu] W. Li and J.H. Rubinstein, *Casson invariant is a homotopy invariant*, Preprint (2002).

[LWu] J. Luecke and Y-Q. Wu, *Relative Euler number and finite covers of graph manifolds*, Geometric topology, (ed. H.Kazez), AMS/, vol. 1, International Press, 1997, 80–103.

[Mc] D. McCullough, *Mapping of reducible 3-manifolds*, Proc. of the Semester in Geometric and Algebraic Topology, Warsaw, Banach Center, Publ., 1986, 61–76.

[N] W.D. Neumann, *Commensurability and virtual fibration for graph manifolds*, Topology **36** (1997), 355–378.

[PW] L. Potyagailo, and S.C. Wang, *3-manifolds with co-Hopf fundamental groups*, St. Petersburg Math. J. (English Translation) **11** (2000), 861–881.

[Re1] A. Reznikov, *Volume of discrete groups and topological complexity of homology spheres*, Math. Ann. **306** (1996), 547–554.

[Re2] ———, *Analytic Topology*, Progress in Math. Vol.201, 519–532. Birkhauser 2002.

[RW] A. Reid and S.C.Wang, *Non-Haken 3-manifolds are not large with respect to mappings of non-zero degree*, Comm. Anal. Geom. **7** (1999), 105–132.

[RWZh] A.Reid, S.C.Wang, Q.Zhou, *Generalized Hopfian property, a minimal Haken manifold, and epimorphisms between 3-manifold groups*, Acta Math. Sinica **18** (2002), 157–172.

[Ro1] Y. Rong, *Degree one maps between geometric 3-manifolds*, Trans. AMS (1992).

[Ro2] ———, *Maps between Seifert fibered spaces of infinite* π_1, Pacific J. Math. **160** (1993), 143–154.

[Ro3] ———, *Degree one maps of Seifert manifolds and a note on Seifert volume*, Topology Appl. **64** (1995), 191–200.

[RoW] Y. Rong and S.C.Wang, *The preimage of submanifolds*, Math. Proc. Camb. Phil. Soc. **112** (1992), 271–279.

[Se1] Z. Sela, *Structure and rigidity in (Gromov) hyperbolic groups*, GAFA **7** (1997), 561–593.

[SWZ] A. Shastri, J.G.Williams and P.Zvengrowski, *Kinks in general relativity*, Internat. J. Theor. Phys. **19** (1980), 1–23.

[So1] T. Soma, *A rigidity theorem for Haken manifolds*, Math. Proc. Camb. Phil. Soc. **118** (1995), 141–160.

[So2] ———, *Non-zero degree maps onto hyperbolic* 3*-manifolds*, J. Diff. Geom. **49** (1998), 517–546.

[So3] ———, *Sequence of degree-one maps between geometric 3-manifolds.*, Math. Ann. **49** (2000), 733–742.

[So4] ———, *Degree one maps between hyperbolic 3-manifolds with the same limit*, Trans. AMS **353** (2001), 2753.

[So5] ———, *Epimorphisms sequences between hyperbolic 3-manifold groups*, Proc. AMS **130** (2002), 1221–1223.

[Th1] W. Thurston, *The Geometry and Topology of Three-Manifolds*, Princeton Lecture Notes.

[Th2] ———, *Three dimensional manifolds, Kleinian groups and hyperbolic geometry*, Bull. AMS **6** (1982), 357–388.

[Th3] ———, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. AMS **19** (1988), 417–431.

[W1] S.C.Wang, *The existence of non-zero degree maps between aspherical 3-manifolds*, Math. Zeit. **208** (1991), 147–160.

[W2] ———, *The π_1-injectivity of self-maps of non-zero degree on 3-manifolds*, Math. Ann. **297** (1993), 171–189.

[WWu1] S.C.Wang and Y.Q.Wu, *Any knot complement covers at most one knot complement*, Pacific J. Math. **158** (1993), 387–395.

[WWu2] ———, *Covering invariant of graph manifolds and cohopficity of 3-manifold groups*, Proc. London Math. Soc. **68** (1994), 221–242.

[WZh1] S.C.Wang and Q.Zhou, *Embeddings of Kleinian groups with torsion*, Acta Math. Sinica **17** (2001), 21–34.

[WZh2] ———, *Any 3-manifold 1-dominates at most finitely many geometric 3-manifolds*, Math. Ann. **332** (2002), 525–535.

[YW] F.Yu and S.C.Wang, *Covering degrees are uniquely determined by graph manifolds involved*, Comm. Math. Helv. **74** (1999), 238–247.

Section 6. Algebraic and Complex Geometry

Hélène Esnault: *Characteristic Classes of Flat Bundles and Determinant of the Gauss-Manin Connection* 471

L. Göttsche: *Hilbert Schemes of Points on Surfaces* 483

Shigeru Mukai: *Vector Bundles on a K3 Surface* 495

R. Pandharipande: *Three Questions in Gromov-Witten Theory* 503

Miles Reid: *Update on 3-folds* 513

Vadim Schechtman: *Sur les Algèbres Vertex Attachées aux Variétés Algébriques* 525

B. Totaro: *Topology of Singular Algebraic Varieties* 533

ICM 2002 · Vol. II · 471–481

Characteristic Classes of Flat Bundles and Determinant of the Gauss-Manin Connection

Hélène Esnault*

2000 Mathematics Subject Classification: 14C22, 14C25, 14C40, 14C35, 14C99.

1. Introduction

The purpose of this note is to give a survey on recent progress on characteristic classes of flat bundles, and how they behave in a family.

2. Characteristic classes

Let X be a smooth algebraic variety over a field k. In [13] and [15], we defined the ring

$$\begin{aligned} AD(X) &= \oplus_n AD^n(X) \\ &= \oplus_n \mathbb{H}^n(X, \mathcal{K}_n^M \xrightarrow{d\log} \Omega^n_{X/k} \xrightarrow{d} \ldots \to \Omega^{2n-1}_{X/k}) \end{aligned} \tag{2.1}$$

of algebraic differential characters. Here the Zariski sheaf $\mathcal{K}_n^M$ is the kernel of the residue map from Milnor K-theory at the generic point of X to Milnor K-theory at codimension 1 points. More precisely, $\mathcal{K}_n^M$ satisfies a Gersten type resolution (see [16] and [18])

$$\begin{aligned} \mathcal{K}_n^M \xrightarrow{\cong} & (i_{k(X),*}K_n^M(k(X)) \xrightarrow{\text{Res}} \oplus_{x\in X^{(1)}} i_{x,*}K^M_{n-1}(\kappa(x)) \to \\ & \ldots \oplus_{x\in X^{(a)}} i_{x,*}K^M_{n-a}(\kappa(x)) \to \ldots \to \oplus_{x\in X^{(n)}} i_{x,*}K^M_0(\kappa(x))). \end{aligned}$$

Here $X^{(a)}$ means the free group on points in codimension a, while $i_x : x \to X$ is the embedding. The map $d\log(\{a_1, \ldots, a_n\}) = d\log a_1 \wedge \cdots \wedge d\log a_n$ from $K_n^M(k(X))$

*Mathematik, Universität Essen, FB6, Mathematik, 45117 Essen, Germany. E-mail: esnault@uni-essen.de

to $\Omega^n_{k(X)/k}$ carries $\mathcal{K}^M_n$ to

$$\Omega^n_{X/k} = \mathrm{Ker}(\Omega^n_{k(X)} \xrightarrow{\mathrm{res}} \oplus_{x\in X^{(1)}}\Omega^{n-1}_{x/k}).$$

This defines the map $d\log : \mathcal{K}^M_n \to \Omega^n_{X/k}$.

By the Gersten resolution, $H^n(X, \mathcal{K}^M_n) = CH^n(X)$, the Chow group of codimension n points. Thus one has a forgetful map

$$AD^n(X) \xrightarrow{\mathrm{forget}} CH^n(X). \tag{2.2}$$

The restriction map to the generic point $\mathrm{Spec}(k(X))$ fulfills

$$\begin{aligned} AD^1(X) &\xrightarrow{\cong} H^0(X, \Omega^1_{X/k}/d\log \mathcal{O}^\times_X) \subset AD^1(k(X)) \\ AD^n(X) &\to H^0(X, \Omega^{2n-1}_{X/k}/d\Omega^{2n-2}_{X/k}) \subset \Omega^{2n-1}_{k(X)/k}/d\Omega^{2n-2}_{k(X)/k}, \quad \text{for } n \geq 2 \end{aligned} \tag{2.3}$$

(see [2]). It is no longer injective for $n \geq 2$.

The Kähler differential $d : \Omega^{2n-1}_{X/k} \to \Omega^{2n}_{X/k,\mathrm{clsd}}$ defines

$$AD^n(X) \xrightarrow{d} H^0(X, \Omega^{2n}_{X/k,\mathrm{clsd}}). \tag{2.4}$$

The ring $AD(X) = \oplus_n AD^n(X)$ contains the subring

$$\begin{aligned} AD(X)_{\mathrm{clsd}}(X) &= \oplus_n AD^n_{\mathrm{clsd}}(X) \oplus_n \mathbb{H}^n(X, \mathcal{K}^M_n \xrightarrow{d\log} \Omega^n_{X/k} \to \ldots \to \Omega^{\dim(X)}_{X/k}) \\ &= \mathrm{Ker}(AD(X) \xrightarrow{d} \oplus_n H^0(X, \Omega^{2n}_{X/k,\mathrm{clsd}})). \end{aligned} \tag{2.5}$$

We call them the closed characters. The restriction map to $\mathrm{Spec}(k(X))$ fulfills

$$\begin{aligned} AD^1(X)_{\mathrm{clsd}} &\xrightarrow{\cong} H^0(X, \Omega^1_{X/k,\mathrm{clsd}}/d\log\mathcal{O}^\times_X) \subset AD^1(k(X))_{\mathrm{clsd}} \\ AD^n(X)_{\mathrm{clsd}} &\to H^0(X, \mathcal{H}^{2n-1}_{DR}) \subset H^{2n-1}_{DR}(k(X)/k), \quad \text{for } n \geq 2 \end{aligned} \tag{2.6}$$

(see [2]). Here $\mathcal{H}^p_{DR}$ is the Zariski sheaf of de Rham cohomology.

If k is the field of complex numbers $\mathbb{C}$, one can change from the Zariski topology to the analytic one. This yields a map

$$\begin{aligned} AD(X) &= \oplus_n AD^n(X) \xrightarrow{l} \\ D(X) &= \oplus_n D^n(X) = \oplus_n \mathbb{H}^{2n}(X_{\mathrm{an}}, \mathbb{Z}(n) \to \mathcal{A}^0_X \xrightarrow{d} \ldots \to \mathcal{A}^{2n-1}_X). \end{aligned} \tag{2.7}$$

Here $D(X)$ is the ring of differential characters defined by Cheeger-Simons ([10]). One has

$$\iota(AD(X)_{\mathrm{clsd}}) \subset \oplus_n H^{2n-1}(X_{\mathrm{an}}, \mathbb{C}/\mathbb{Z}(n)) \subset D(X). \tag{2.8}$$

It is classical that $AD^1(X)$ is the group of isomorphism classes of lines bundles L with connection ∇. If $\xi_{ij} \in \Gamma(U_{ij}, \mathcal{O}^\times_X)$ is a cocycle of L in a local frame e_i of

L on U_i, and $\nabla(e_i) = \alpha_i \in \Gamma(U_i, \Omega^1_{X/k})$ is the local form of the connection, then $d\log \xi_{ij} = \alpha_j - \alpha_i = \delta(\alpha)_{ij}$ defines the Cech cocycle of $c_1((L,\nabla))$. In [12], [13], [15], we generalize this class.

Theorem 2.1 ([12], [13], [15]). *Associated to an algebraic bundle E with connection ∇ (resp. with integrable connection), one has characteristic classes $c_n((E,\nabla)) \in AD^n(X)$ (resp. $\in AD^n(X)_{\mathrm{clsd}}$). These classes satisfy the following properties:*

(1) *The classes $c_*((E,\nabla)) \in AD^*(X)$ are functorial and additive.*
(2) *$c_1((E,\nabla))$ is the isomorphism class of $(\det(E), \det(\nabla))$.*
(3) *forget $(c_n((E,\nabla))) = c_n(E) \in CH^n(X)$ is the Chern class of the underlying algebraic bundle E in the Chow group.*
(4) *$d(c_n((E,\nabla))) = c_n(\nabla^2) \in H^0(X, \Omega^{2n}_{X/k,\mathrm{clsd}})$ is the Chern-Weil form which is the evaluation of the invariant polynomial c_n on the curvature ∇^2.*
(5) *The restriction to the generic point $c_n((E,\nabla)|_{k(X)})$ is the algebraic Chern-Simons invariant $CS_n((E,\nabla))$ defined in [4]. It has values in $H^0(X, \Omega^1_{X/k}/d\log \mathcal{O}^\times)$ $(resp. H^0(X, \Omega^1_{X/k,\mathrm{clsd}}/d\log \mathcal{O}^\times))$ for $n = 1$, and in $H^0(X, \Omega^{2n-1}_{X/k}/d\Omega^{2n-2}_{X/k})$ $(resp. H^0(X, \mathcal{H}^{2n-1}_{DR}))$ for $n \geq 2$.*
(6) *If $k \subset \mathbb{C}$, then $\iota(C_n((E,\nabla))) \in D^n(X) (resp. \in H^{2n-1}(X_{\mathrm{an}}, \mathbb{C}/\mathbb{Z}(n)))$ is the differential character defined by Chern-Cheeger-Simons, denoted by $c_n(E^\nabla_{\mathrm{an}}) \in H^{2n-1}(X_{\mathrm{an}}, \mathbb{C}/\mathbb{Z}(n))$ if ∇ is integrable.*

If $k = \mathbb{C}, \nabla$ is flat and the underlying monodromy is finite, then the existence of $c_n((E,\nabla))$immediately implies that the Chern-Simons classes in $H^{2n-1}(X_{\mathrm{an}}, \mathbb{Q}(n)/\mathbb{Z}(n))$ are in the smallest possible level of the coniveau filtration ([14]).

If X is complex projective smooth, ∇ is integrable and $n \geq 2$, we relate $CS_n((E,\nabla))$ for $n \geq 2$ to the (generalized) Griffiths' group $\mathrm{Griff}^n(X)$. It consists of cycles which are homologous to 0 modulo those which are homologous to 0 on some divisor([4], definition 5.1.1). For $n = 2$, [2] implies that $\mathrm{Griff}^2(X)$ is the classical Griffiths' group. For $n \geq 2$, Reznikov's theorem ([19]) (answering positively Bloch's conjecture [3]), together with the existence of the lifting $c_n((E,\nabla))$, imply that the classes $CS_n((E,\nabla))$ lie in the image Im of the global cohomology $H^{2n-1}(X, \mathbb{Q}(n))$ in $H^0(X, \mathcal{H}^{2n}(\mathbb{Q}(n))$. This subgroup Im maps to Griff^n(X). One has

Theorem 2.2 ([4], Theorem 5.6.2).

$$\text{image of } CS_n((E,\nabla)) \in \mathrm{Griff}^n(X) \otimes \mathbb{Q}$$

is the Chern class $c_n^{\mathrm{Griff}}(E) \otimes \mathbb{Q}$ of the underlying algebraic bundle E. Moreover, $CS_n((E,\nabla)) = 0$ if and only if $c_n^{\mathrm{Griff}}(E) \otimes \mathbb{Q} = 0$.

A relative version $AD(X/S)$ of $AD(X)$ is defined in [6]. We give here an example of application.

Theorem 2.3([6], Corollary 3.15). *Let $f : X \to S$ be a smooth projective family of curves over a field k. Let $(E, \nabla_{X/S})$ be a bundle with a relative connection. Then there are classes $c_2((E,\nabla)) \in AD^2(X/S) := \mathbb{H}^2(X, \mathcal{K}_2 \xrightarrow{d\log} \Omega^1_S \otimes^1_{X/S})$ lifting the classes $c_2(E) \in CH^2(X)$. There is a trace map $f_* : AD^2(X/S) \to AD^1(S)$ compatible with the trace map on Chow groups $f_* : CH^2(X) \to CH^1(S)$. Thus*

$f_*c_2(E,\nabla)$ is a connection on the line bundle $f_*c_2(E)$, which depends functorially on the choice of $\nabla_{X/S}$ on E.

We now study the behavior at ∞ of $CS_n((E,\nabla))$ for $n \geq 2$ in characteristic 0. Let $j : X \to \bar{X}$ be a smooth compactification of X. Recall that a de Rham class $\in H^q_{DR}(k(X)/k)$ at the generic point is called *unramified* if it lies in $H^0(\bar{X}, \mathcal{H}^q_{DR} \subset H^q_{DR}(k(X)/k)([2])$.

Theorem 2.4. *We assume k to be of characteristic 0, and ∇ to be integrable. Then $CS_n((E,\nabla))$ is unramified for $n \geq 2$.*

Proof. If (E,∇) is regular singular, this is shown in [4], theorem 6.1.1. In general, one may argue as follows. One has

$$H^0(\bar{X}, \mathcal{H}^{2n-1}_{DR}) = \mathrm{Ker}(H^0(X, \mathcal{H}^{2n-1}_{DR}) \xrightarrow{\mathrm{Res}} \oplus_{x \in X^{(1)}} H^{2n-2}_{DR}(x/k)).$$

Thus it is enough to show that the residue map at each generic point at ∞ dies. At a smooth point of a divisor D at ∞, the residue depends only on the formal completion of X along D. So we may assume that ∇ is a connection on $\mathcal{O}_X = k(D)[[x]]$, integrable relative to k. By a variant (see [1], proposition 5.10.) of Levelt's theorem ([17]) for absolute flat connections, there are a finite extension $K \supset k(D)$, and a ramified extension $\pi : K[[x]] \subset K[[y]], y^N = x$ for some $N \in \mathbb{N}\backslash\{0\}$ such that $\pi^*(\nabla) = \oplus(L \otimes U)$. Here L is integrable of rank 1 and U is integrable with logarithmic poles along y=0. Since $\mathrm{Res}_{y=0}\pi^*(\alpha) = N\pi^*(\mathrm{Res}_{x=0}(\alpha))$, and $H^p_{DR}(k(D)/k) \subset H^p_{DR}(K/k)$, functoriality and additivity of the classes imply that we may assume $\nabla = L \otimes U$ on $K((x))$ with $K = k(D)$, In a local frame we have the equations $U = \Gamma\dfrac{dx}{x} + \Sigma$ where $\Gamma \in GL(r, K[[x]]), \Sigma \in M(r, K[[x]]) \otimes \Omega^1_K$, and $L = d(f) + \lambda\dfrac{dx}{x} + \beta$ where $f \in K((x)), \lambda \in k, \beta \in \Omega^1_K$ with $d\beta = 0$. The explicit formula Res $\mathrm{Tr}(A(d(A))^{n-1}) \in H^{2n-1}_{DR}(K)$ of $CS_n((E,\nabla)) \in H^{2n-1}_{DR}(K((x)))$ and [4], prop. 5.10 in the logarithmic case, imply that

$$CS_n((E,\nabla)) = \mathrm{Tr}(d(f) + \lambda\frac{dx}{x} + \beta)(d(\Gamma)\frac{dx}{x} + d\Sigma)^{n-1}.$$

This is the sum of 2 terms with rational coefficients, Res $\mathrm{Tr}(d(f) + \beta)(d(\Sigma)^{n-2}$ $d(\Gamma)\dfrac{dx}{x})$ and Res $\mathrm{Tr}(\lambda\dfrac{dx}{x}(d(\Sigma)^{n-1}))$. Both terms are obviously exact.

Discussion 2.5. We assume here $k = \mathbb{C}$, ∇ is integrable and $n \geq 2$. We consider the image $c_n(E^\nabla_{\mathrm{an}}|_{\mathbb{C}(X)}) \in H^0(\bar{X}, \mathcal{H}^{2n-1}(\mathbb{C}/\mathbb{Z}(n)))$ of $CS_n((E,\nabla)) \in H^0(\bar{X}, \mathcal{H}^{2n-1}_{DR})$. When X is not compact, there is Deligne's unique algebraic (E,∇) with regular singularities at ∞ with the given underlying local system E^∇_{an}([11]), but there are many irregular connections (E,∇). The topological class $C_n(E^\nabla_{\mathrm{an}}) \in H^{2n-1}(X_{\mathrm{an}}, \mathbb{C}/\mathbb{Z}(n))$ is not, a priori, extendable to $\bar{X}$, but we have seen that its restriction to $\mathrm{Spec}(\mathbb{C}(X))$ is unramified.

There is on X a fundamental system of Artin neighborhoods U which are geometrically successive fiberings in affine curves. Topologically they are $K\pi_1$ and their fundamental group is a successive extension of free groups in finitely many letters. On such an open U, the class $c_n(E^\nabla_{\mathrm{an}})$ lies in $H^{2n-1}(U_{\mathrm{an}}, \mathbb{C}/\mathbb{Z}(n)) = H^{2n-1}(\pi_1(U_{\mathrm{an}}, u), \mathbb{C}/\mathbb{Z}(n))$.

If U is such that $\pi_1(U_{\mathrm{an}}, u)$ is isomorphic as an abstract group to $\pi_1(V_{\mathrm{an}}, v)$, where V is an Artin neighborhood on a rational variety, then $E^{\nabla}_{\mathrm{an}}|_U$ becomes a representation of $\pi_1(V_{\mathrm{an}}, v)$, and since then $H^0(\bar{V}, \mathcal{H}^{2n-1}_{DR}) = 0$ for $n \geq 2$ and $V \subset \bar{V}$ a good compactification, one obtains $c_n(E^{\nabla}_{\mathrm{an}}|_{\mathbb{C}(X)}) = 0, n \geq 2$ in this case. Such an example is provided by a product of smooth affine curves of any genus. It has the fundamental group of a product of $\mathbb{P}^1$ minus finitely many points.

Question 2.6. In view of the previous discussion, we may ask what complex smooth varieties X are dominated by $h : Y \to X$ proper, with Y smooth, such that Y has an Artin neighborhood, the fundamental group of which is the fundamental group of an Artin neighborhood on a rational variety, or more generally of a variety for which $H^0(\text{smooth compactification}, \mathcal{H}^n_{DR}) = 0$ for $n \geq 1$. We have seen that this would imply vanishing modulo torsion of $c_n(E^{\nabla}_{\mathrm{an}}|_{\mathbb{C}(X)})$, $n \geq 2$, or equivalently $CS_n((E, \nabla)) \in H^0(\bar{X}, \mathcal{H}^{2n-1}(\mathbb{Q}(n)))$.

On the other hand, if X is projective smooth, Reznikov's theorem ([19]) shows vanishing modulo torsion of the Chern-Cheeger-Simons classes in $H^{2n-1}(X_{\mathrm{an}}, \mathbb{C}/\mathbb{Q}(n))$. It is a consequence of Simpson's nonabelian Hodge theory on smooth projective varieties. Our classes $CS_n((E, \nabla))$ live at the generic point of X. We don't have a nonabelian mixed Hodge theory at disposal. Yet one may ask whether it is always true that $CS_n((E, \nabla)) \in H^0(\bar{X}, \mathcal{H}^{2n-1}(\mathbb{Q}(n)))$ for $n \geq 2$, even if many X don't have the topological property explained above.

3. The behavior of the algebraic Chern-Simons classes in families in the regular singular case

The algebraic Chern-Simons invariants $CS_n((E, \nabla))$ have been studied in a family in [5]. Given $f : X \to S$ a proper smooth family, and (E, ∇) a flat connection on X, the Gauß-Manin bundles

$$R^i f_*(\Omega^{\bullet}_{X/S} \otimes E, \nabla_{X/S})$$

carry the Gauß-Manin connection $GM^i(\nabla)$. We give a formula for the invariants $CS_n((GM^i(\nabla) - \text{rank } (\nabla) \cdot GM^i(d)))$ on S, as a function of $CS_n((E, \nabla))$ and of characteristic classes of f. Here (O, d) is the trivial connection.

More generally, we may assume that f is smooth away from a normal crossings divisor $T \subset S$ such that $Y = f^{-1}(T) \subset X$ is a normal crossings divisor with the property that $\Omega^1_{X/S}(\log Y)$ is locally free. Then (E, ∇) has logarithmic poles along $Y \cup Z$ where $Y \cup Z \subset X$ is a normal crossings divisor, still with the property that $\Omega^1_{X/S}(\log(Y + Z))$ is locally free. That is Z is the horizontal divisor of singularities of ∇. The formula involves the top Chern class $c_d(\Omega^1_{X/S}(\log(Y + Z)) \in \mathbb{H}^d(X, \mathcal{K}_d \to \oplus_i \mathcal{K}_{Z_{i,d}})$, rigidified by the residue maps $\Omega^1_{X/S}(\log(Y + Z)) \to \mathcal{O}_{Z_i}$, as defined by T. Saito in [20]. One of its main features is that $CS_n((GM^i(\nabla) - \text{rank } (\nabla) \cdot GM^i(d)))$ vanishes if $CS_n((E, \nabla))$ vanishes. It is

Theorem 3.1 ([5], Theorem 0.1).

$$\begin{aligned} & CS_n(\sum(-1)^i(GM^i(\nabla) - \mathrm{rank}(\nabla)\cdot GM^i(d))) \\ = \; & (-1)^{\dim(X/S)} f_* c_{\dim(X/S)}(\Omega^1_{X/S}(\log(Y+Z)), \mathrm{res}) \cdot CS_n((E,\nabla)). \end{aligned}$$

Here $\cdot$ is the cup product of the algebraic Chern-Simons invariants with this rigidified class, which is well defined, as well as the trace f_* to S.

Discussion 3.2. One weak point of the method used in [5] is that it does not allow to understand a formula for the whole invariants $c_n((E,\nabla))$, but only for $CS_n(E,\nabla)$. Indeed, we use the explicit formula studied in [4] to compute it, which can't exist for the whole class in $AD(X)$, as it in particular involves the Chern classes of the underlying algebraic bundle E in the Chow group.

4. The determinant of the Gauss-Manin connection: the irregular rank 1 case

Now we no longer assume that (E,∇) is regular singular at ∞. In the next two sections, we reduce ourselves to the case where $f : X \to S$ is a family of curves, and we consider only the determinant of the Gauß-Manin connection. That is we consider

$$\begin{aligned} \det(GM) \; &:= \; \sum_i (-1)^i c_1(GM^i) \\ &\in \; \mathbb{H}^1(S, \mathcal{O}_S^\times \xrightarrow{d\log} \Omega^1_S) \subset \Omega^1_{k(S)}/d\log(k(S))^\times. \end{aligned}$$

Since the determinant is recognized at the generic point of S, we replace S by its function field $K := k(S)$ in the next two sections. In other words, X/K is an affine curve. Let $\bar{X}/K$ be its smooth compactification.

In this section, we assume that the integrable connection (L,∇) we start with on X has rank 1. Following Deligne's idea (see his 1974 letter to Serre published as an appendix to [7], we first reduce the problem of computing the determinant of the cohomology of ∇ on the curve to the one of computing the determinant of cohomology of an integrable invariant connection, still denoted by ∇, on a generalized Jacobian. More precisely, ∇ has a divisor (with multiplicities) of irregularity $\sum_i m_i p_i$, where $m_i - 1$=irregularity of ∇ in $p_i \in \bar{X} \setminus X$. On the Jacobian $G = \mathrm{Pic}(\bar{X}, \sum_i m_i p_i)$ of line bundles trivialized at the order m_i at the points p_i, there is an invariant connection which pulls back to ∇ via the cycle map. On the torsor $p^{-1}(\omega_{\bar{X}}(\sum_i m_i p_i))$ under the affine group $p^{-1}(\mathcal{O}_{\bar{X}})$, where $p : G \to \mathrm{Pic}(\bar{X})$, one considers the hypersurface $\Sigma : \sum_i \mathrm{res}_{p_i} = 0$. We show that the relative invariant connection $\nabla_{/K}$ on G, while restricted to $\Sigma \subset p^{-1}(\omega_{\bar{X}}(\sum_i m_i p_i))$, acquires exactly one zero which is a K-rational point κ of G. Restricting ∇ to this special point yields a connection $\nabla|_\kappa$ on K. The formula then says that the determinant of the Gauß-Manin connection is the sum of this connection $\nabla|_\kappa$ and of a 2-torsion term, which we describe now. In a given frame of L at a singularity p_i, the local equation

of the connection is $\alpha_i = a_i \frac{dt_i}{t_i^{m_i}}$+lower order terms, where t_i is a local parameter and $a_i \in k^\times$. Then the 2-torsion connection $\frac{m_i}{2} d\log a_i \in \Omega^1_{K/k}/d\log K^\times$ does not depend on the choices and is well defined. The 2-torsion term is the sum over the irregular points of these 2-torsion connections. Summarizing, one has

Theorem 4.1 ([7], Theorem 1.1).

$$\begin{aligned} &\det\left(\sum_i (-1)^i H^i(X, (\Omega^\bullet_{X/K} \otimes L, \nabla_{X/K}), GM^i(\nabla)\right) \\ = \;& (-1)\nabla|_\kappa + \sum_{i, m_i \geq 2} \frac{m_i}{2} d\log a_i \in \Omega^1_{K/k}/d\log K^\times. \end{aligned}$$

Discussion 4.2. The formula described above is global. As such, it has a spirit which is different from Deligne's formula describing the global ϵ-factor of an ℓ-adic character over a curve over a finite field as a product of local ϵ-factors. However, choosing another K-rational point $\kappa' \in p^{-1}(\omega_{\bar{X}}(\sum_i m_i p_i))$, it is easy to write the difference $\nabla|_{\kappa'} - \nabla|_\kappa$ as a sum of explicitly given connections on K. One has $u \cdot \kappa' = \kappa$, where $u = \prod u_i \in p^{-1}\mathcal{O}_{\bar{X}} = \prod_i (\mathcal{O}_{\bar{X}, p_i}/\mathfrak{m}_{\mathfrak{p}_i})^\times / K^\times$. Then $\nabla_{\kappa'} = \nabla|_\kappa - \sum_i \mathrm{Res}_{p_i} d\log u_i \wedge \alpha_i$. Correspondingly, one may write the right hand side of the formula above as

$$(-1)\nabla|_{\kappa'} + \sum_i \left(\sup\left(1, \frac{m_i}{2}\right) d\log a_i + \mathrm{Res}_{p_i} d\log u_i \wedge \alpha_i\right).$$

In particular, the choice of some differential form $\nu \in \omega_{\bar{X}/K} \otimes K(X)$, generating $\omega_{\bar{X}/K}(\sum m_i p_i)$ at the point p_i, defines a trivialization of $\omega_{\bar{X}/K}(\sum_i m_i p_i)$ thus a point $\kappa(\nu)$. We write $\alpha_i = g_i \nu$ with $g_i \in (\mathcal{O}_{X, p_i}/\mathfrak{m}_{\mathfrak{p}_i})^\times$. Then the formula reads

Theorem 4.3 ([7], Formula 5.4).

$$\begin{aligned} &\det\left(\sum_i (-1)^i (H^i(X, (\Omega^\bullet_{X/K} \otimes L, \nabla_{X/K}), GM^i(\nabla)\right) \\ = \;& (-1)\det(\nabla)|_{\kappa(\nu)} + \sum_i \left(\sup\left(1, \frac{m_i}{2}\right) d\log a_i + \mathrm{Res}_{p_i} dg_i g_i^{-1} \wedge \alpha_i\right). \end{aligned}$$

5. The determinant of the Gauss-Manin connection: the irregular higher rank case

We assume in this section that we have an affine curve X over $K = k(S)$, k of characteristic zero as in the rank 1 case. The integrable connection (E, ∇) we are given on X has higher rank r.

In the rank one case, for any rank one bundle contained in $j_* E$, the equation of the connection in a local formal frame at a singular point is of the shape $\alpha = a\frac{dt}{t^m} + \frac{\beta}{t^{m-1}}$, where $m \in \mathbb{N}, a \in K[[t]]^\times$, $\beta \in \Omega^1_K \otimes K[[t]]$. In particular, $(m-1)$ is the irregularity of the connection ([11]). Here t is a local parameter. If $r > 1$, it is

no longer the case that j_*E necessarily contains a rank r bundle such that in a local formal frame of this bundle, the local equation has the shape $A_i = a_i \frac{dt_i}{t_i^m} + \frac{\beta_i}{t_i^{m-1}}$, with $a_i \in GL(r, K[[t_i]]), \beta_i \in \Omega^1_K \otimes M(r, K[[t_i]])$. We call an integrable connection (E, ∇) with this existence property an *admissible* connection.

Even if (E, ∇) is admissible, its determinant connection $\det(E, \nabla)$ might have much lower order poles (for example trivial). This indicates that one can not extend directly in this form the formula 4.1. However, assuming (E, ∇) to be admissible and choosing some $\nu \in \omega_{\bar{X}/K} \otimes K(X)$ which generates $\omega(\sum_i m_i p_i)$ at p_i as for formula 4.3, the right hand side of 4.3 makes sense, if one replaces $d \log a_i$ by $d \log \det(a_i)$. Using global methods inspired by the Higgs correspondence between Higgs fields and connections on complex smooth projective varieties ([21]), one is able to prove the "same" formula as 4.3 in the higher rank case on $\mathbb{P}^1$.

Theorem 5.1 ([8], Theorem 1.3). *If (E, ∇) is admissible and has at least one irregular point, and if $\nu \in \omega_{\bar{X}/K} \otimes K(X)$ generates $\omega(\sum_i m_i p_i)$ at the points p_i, then*

$$\begin{aligned} & \det\left(\sum_i (-1)^i (H^i(X, (\Omega^\bullet_{X/K} \otimes L, \nabla_{X/K}), GM^i(\nabla))\right) \\ = \quad & (-1)\nabla|_{\kappa(\nu)} + \sum_i \left(\sup\left(1, \frac{m_i}{2}\right) d \log \det(a_i(p_i)) + \operatorname{Tr} \operatorname{Res}_{p_i} dg_i g_i^{-1} \wedge A_i\right). \end{aligned}$$

The connection $\operatorname{Res} \operatorname{Tr}_{p_i} dg_i g_i^{-1} \wedge A_i \in \Omega^1_K / d \log K^\times$ is well defined, as well as the 2-torsion connection $\sup\left(1, \frac{m_i}{2}\right) d \log \det(a_i(p_i))$.

However, one needs a different method in order to understand the contribution of singularities in which (E, ∇) is not admissible.

We describe now the origin of the method contained [1]. It is based on the idea that Tate's method ([22]) applies for connections.

Locally formally over the Laurent series field $K((t))$, E becomes a r-dimensional vector space over $K((t))$. The relative connection $\nabla_{K((T))/K} : E \to \omega_{K((t))/K} \otimes E$ is a Fredholm operator. This means that $H^i(\nabla_{X/K}), i = 0, 1$ are finite dimensional K-vector spaces, and that $\nabla_{X/K}$ carries compact lattices to compact lattices. Let $E \cong \oplus_1^r K((t))$ be the choice of a local frame. A compact lattice is a K-subspace of E which is commensurable to $\oplus_1^r K[[t]]$. Given $0 \neq \nu \to \in \omega_{K((t))/K}$, one composes $\nabla_{K((t))/K, \nu} := \nu^{-1} \circ \nabla_{K((t))/K} : E \to E$ to obtain a Fredholm endomorphism. To a Fredholm endomorphism $A : E \to E$, one associates a 1-dimensional K-vector space $\lambda(A) = \det(H^0(A)) \otimes \det(H^1(A))^{-1}$ together with the degree $\chi(A) = \dim H^0(A) - \dim H^1(A)$. We call this a *super-line*. It does not refer to the topology defined by compact lattices. Then one measures how A moves a compact lattice $L \subset E$. First for 2 lattices L and L', one takes a smaller compact lattice $N \subset L \cap L'$ and defines $\det(L : L') := \det(L/N) \cdot \det(L'/N)^{-1}$, where $\cdot$ is the tensor product of super-lines and $\det(L/N)$ has degree $\dim(L/N)$. This does not depend on the choice of N. Then one defines asymptotic superlines. The compact one is $\lambda_c(A) = \det(A(L) : L) \cdot \det(L \cap \operatorname{Ker}(A))$ and the discrete one is $\lambda_d(A) = \det(L : A^{-1}(L)) \cdot \det(V/(L + A(V))$. They do not depend on the choice of

L. Taking $0 \neq \nu \in \omega_{\bar{X}/K} \otimes K(X)$ a rational differential form, and $E_{\min}$ the minimal extension of E, X the complement of the singularities of ν and $\nabla_{X/K}$, one has

$$\det H^*(X/K, E_{\min}) = \otimes_{X \in \bar{X} \setminus X} \lambda_d(\nabla_{K((t))/K,\nu}) \cdot \det(E^{\nabla})^{-1}$$

as a product of discrete lines.

On the other hand, one has the relation $\lambda(A) = \lambda_d(A) \cdot \lambda_c(A)^{-1}$. One easily computes that $\chi(\nabla_{K((t))/K,\nu}) = 0, \lambda(\nabla_{K((t))/K,\nu}) = 1$. Setting

$$\epsilon_x(\nabla_{K((t))/K,\nu}) = \lambda_c(\nabla_{K((t)/K,\nu}) \cdot \det(E^{\nabla})^{-1},$$

this implies immediately the product formula.

Theorem 5.2 ([1], (1.3.1)).

$$\det H^*(X/K, E_{\min}) = \oplus_{x \in \bar{X} \setminus X} \epsilon_x(\nabla_{K((t))/K,\nu}).$$

It remains to endow the local ϵ lines with a connection, compatible with the Gauß-Manin connection on the left. One chooses a section of the vector fields $T_{K/k} \subset T_{X/k}$ and a relative differential form ν which is annihilated by this section. One applies Grothendieck's definition of a connection. The Gauß-Manin connection is given by the infinitesimal automorphism $p_1^* \det H^*(X/K, E_{\min}) \to p_2^* \det H^*(X/K, E_{\min})$ on $K \otimes_k K$, induced by $\tau : p_1^* E_{\min} \to p_2^* E_{\min}, \tau \in T_{K/k}$, which by the choice commutes to $\nabla_{X/K}$. By functoriality of the ϵ lines, this defines a $K \otimes_k K$ homomorphism $p_1^* \epsilon \to p_2^* \epsilon$. This is the ϵ-connection.

Theorem 5.3 ([1], Theorem 5.6). *For an admissible connection of local equation* $A = a\dfrac{dt}{t^m} + \dfrac{\beta}{t^{m-1}}$, *with* $m \geq 2$, *the local* ϵ*-connection is*

$$\begin{aligned} \epsilon\big(\nabla_{K((t))/K}, \frac{dt}{t^m}\big) &= \mathrm{TrRes}_{t=0} daa^{-1}A + \frac{m}{2} d\log\det(a(t=0)) \\ &\in \Omega^1_{K/k}/d\log K^{\times}. \end{aligned}$$

The restriction on the choice of ν given by the commutativity constraint with some lifting of vector fields of K is not necessary. The construction is more general.

Given a relative connection $\nabla_{K((t))/K}$, the ϵ-lines for $0 \neq \nu \in \omega^{\times}_{K((t))/K}$ build a super-line bundle on the ind-scheme $\omega^{\times}_{K((t))/K}$. The line bundle obeys a connection relative to K on $\omega^{\times}_{K((t))/K}$. Formula 5.2 identifies line bundles with connections relative to K, where the left hand side carries the constant connection. The choice of an integrable lifting ∇ of $\nabla_{X/K}$ yields a lifting of the relative connection on the ϵ line to an integrable connection relative to k. Formula 5.2 identifies line bundles with integrable connections where the left hand side carries the Gauß-Manin connection ([1], 1.3).

The ϵ lines and connections are additive in exact sequences and compatible with push-downs. By a variant of Levelt's theorem for integrable formal connections, this allows to show that all connections are induced from admissible ones, for which we have the formula 5.3.

Question 5.4. We don't know how to precisely relate the algebraic group viewpoint developed to treat the rank 1 case, and the special rational point found there, with the polarized Fredholm line method which works in general.

Acknowledgements. I thank the mathematicians I have worked with on the material exposed in those notes. A large part of it has been jointly developed with Spencer Bloch. It is a pleasure to acknowledge the impact of his ideas on a programme I had started earlier and we continued together. I thank Alexander Beilinson. An unpublished manuscript of his and David Kazhdan allowed me to understand completely one of the two constructions explained in [15]. His deep viewpoint reflected in [1] changed the understanding of the formula we had as explained in [8]. I thank Pierre Deligne, whose ideas on epsilon factors have shaped much of my thinking. His letter to Serre on the rank 1 case is published as an appendix to [7], but the content of his seminar at the IHES in 1984 has not been available to me. I thank Takeshi Saito for his willingness to explain different aspects of the ℓ -adic theory.

References

[1] Beilinson A., Bloch S., Esnault H., ε-factors for Gauß-Manin Determinants, preprint 2001, 62 pages.

[2] Bloch S., Ogus A., Gersten's conjecture and the homology of schemes, Ann. Sc. Éc.Normale Sup. **IV**, sér. 7 (1974), 181–201.

[3] Bloch S., Applications of the dilogarithm function in algebraic K-theory and algebraic geometry, Proc. int. Symp. on Alg. Geom., Tokyo 1977, 103–114 (1977).

[4] Bloch S., Esnault H., Algebraic Chern-Simons theory. Am. J. of Mathematics **119** (1997), 903–952.

[5] Bloch S.,Esnault H., A Riemann-Roch theorem for flat bundles, with values in the algebraic Chern-Simons theory, Annals of Mathematics **151** (2000), 1–46.

[6] Bloch S., Esnault H., Relative Algebraic Characters, preprint 1999, 25, appears in the Irvine Lecture Notes.

[7] Bloch S., Esnault H., Gauß-Manin determinants of rank 1 irregular connection on curves, Math.Ann. **321** (2001), 15–87, with an addendum: the letter of P. Deligne to J.-P.Serre (Feb.74) on ϵ-factors, 65–87.

[8] Bloch S., Esnault H., A formula for Gauß-Manin determinants, preprint 2000, 37.

[9] Chern S., Simons J., Characteristic forms and geometric invariants, Ann. of Maths II.ser **99** (1974), 48–69.

[10] Cheeger, J., Simons J., Differential characters and geometric invariants, Geometry and Topology, Proc. Special Year College Park/Md. 1983/1984, Lect. Notes Math. **1167** (1985), 50–80.

[11] Deligne P., Équations Différentielles à Points Singuliers Réguliers, Lect. Notes in Mathematics, **163** (1970), Springer-Verlag.

[12] Esnault H., Characteristic classes of flat bundles. Topology **27** (1988), 323–352.

[13] Esnault H., Characteristic classes of flat bundles, II. K-theory, **6** (1992), 45–56.

[14] Esnault H., Coniveau of Classes of Flat Bundles Trivialized on a Finite Smooth Covering of a Complex Manifold, K-Theory, **8** (1994), 483–497.

[15] Esnault H., Algebraic differential characters, College Park/Md. 1983/1984, Lect. Notes Math. **1167** (1985), 50–80. in Regulators in Analysis, Geometry and Number Theory, Progress in Mathematics, Birkhäuser Verlag, **171** (2000), 89–117.

[16] Kato K., Milnor K-theory and the Chow group of zero cycles, Applications of algebraic K-theory to algebraic geometry and number theory, Proc. AMS-IMS-SIAM Joint Summer Research Conf. Boulder/Colo.(1983), Part 1, Contemp. Math., **55** (1986), 241–253.

[17] Levelt G., Jordan decomposition for a class of singular differential operators, Ark. Math. **13** (1975), 1–27.

[18] Rost M., Chow groups with coefficients, Doc. Math., J. DMV, **1** (1996), 209–214.

[19] Reznikov A., All regulators of flat bundles are torsion, Ann. Math., (2) **141** (1995), 373–386.

[20] Saito T., ϵ-factor of a tamely ramified sheaf on a variety, Inventiones Math. **113** (1993), 389–417.

[21] Simpson C., Higgs bundles and local systerms, Inst. Hautes Études Sci. Publ. Math., **75** (1992), 5–95.

[22] Tate J., Residues of differentials on curves, Ann. Sci. École Norm. Sup., sér. 4, **1** (1968), 149–159.

ICM 2002 · Vol. II · 483–494

Hilbert Schemes of Points on Surfaces

L. Göttsche*

Abstract

The Hilbert scheme $S^{[n]}$ of points on an algebraic surface S is a simple example of a moduli space and also a nice (crepant) resolution of singularities of the symmetric power $S^{(n)}$. For many phenomena expected for moduli spaces and nice resolutions of singular varieties it is a model case. Hilbert schemes of points have connections to several fields of mathematics, including moduli spaces of sheaves, Donaldson invariants, enumerative geometry of curves, infinite dimensional Lie algebras and vertex algebras and also to theoretical physics. This talk will try to give an overview over these connections.

2000 Mathematics Subject Classification: 14C05, 14J15, 14N35, 14J80.
Keywords and Phrases: Hilbert scheme, Moduli spaces, Vertex algebras, Orbifolds.

0. Introduction

The Hilbert scheme $S^{[n]}$ of points on a complex projective algebraic surface S is a a parameter variety for finite subschemes of length n on S. It is a nice (crepant) resolution of singularities of the n-fold symmetric power $S^{(n)}$ of S. If S is a K3 surface or an abelian surface, then $S^{[n]}$ is a compact, holomorphic symplectic (thus hyperkähler) manifold. Thus $S^{[n]}$ is at the same time a basic example of a moduli space and an example of a nice resolution of singularities of a singular variety. There are a number of conjectures and general phenomena, many of which originating from theoretical physics, both about moduli spaces for objects on surfaces and about nice resolutions of singularities. In all of these the Hilbert scheme of points can be viewed as a model case and sometimes as the main motivating example. Hilbert schemes of points on a surface have connections to many topics in mathematics, including moduli spaces of sheaves and vector bundles, Donaldson invariants, Gromov-Witten invariants and enumerative geometry of curves, infinite dimensional Lie algebras and vertex algebras, noncommutative geometry and also theoretical physics.

*Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34014 Trieste, Italy. E-mail: gottsche@ictp.trieste.it

It is usually best to look at the Hilbert schemes $S^{[n]}$ for all n at the same time, and to study their invariants in terms of generating functions, because new structures emerge this way. For Euler numbers, Betti numbers and conjecturally for the elliptic genus these generating functions will be modular forms and Jacobi forms. This fits into general conjectures from physics about invariants of moduli spaces. Also the cohomology rings of the $S^{[n]}$ for different n are closely tied together. The direct sum over n of all the cohomologies is a representation for the Heisenberg algebra modeled on the cohomology of S, and the cohomology rings of the $S^{[n]}$ can be described in terms of vertex operators. In the case that the canonical divisor of the surface S is trivial, this leads to an elementary description of the cohomology rings of the $S^{[n]}$, which coincides with the orbifold cohomology ring of the symmetric power, giving a nontrivial check of a conjecture relating the cohomology ring of a nice resolution of an orbifold to the recently defined orbifold cohomology ring.

The Hilbert schemes $S^{[n]}$ are closely related to other moduli spaces of objects on S, including moduli of vector bundles and moduli of curves e.g. via the Serre correspondence and the Mukai Fourier transform. This leads to applications to the geometry and topology of these moduli spaces, to Donaldson invariants, and also to formulas in the enumerative geometry of curves on surfaces and Gromov-Witten invariants. We want to explain some of these results and connections. We will not attempt to give a complete overview, but rather give a glimpse of some of the more striking results.

1. The Hilbert scheme of points

In this article S will usually be a smooth projective surface over the complex numbers. We will study the Hilbert scheme $S^{[n]} = Hilb^n(S)$ of subschemes of length n on S. The points of $S^{[n]}$ correspond to finite subschemes $W \subset S$ of length n, in particular a general point corresponds just to a set of n distinct points on S. $S^{[n]}$ is projective and comes with a universal family $Z_n(S) \subset S^{[n]} \times S$, consisting of the (W, x) with $x \in W$. An important role in applications of $S^{[n]}$ is played by the tautological vector bundles $L^{[n]} := \pi_* q^*(L)$ of rank n on $S^{[n]}$. Here $\pi : Z_n(S) \to S^{[n]}$ and $q : Z_n(S) \to S$ are the projections and L is a line bundle on S.

Closely related to $S^{[n]}$ is the symmetric power $S^{(n)} = S^n/G_n$, the quotient of S^n by the action of the symmetric group G_n. The points of $S^{(n)}$ correspond to effective 0-cycles $\sum n_i[x_i]$, where the x_i are distinct points of S and the sum of the n_i is n. The forgetful map

$$\rho : S^{[n]} \to S^{(n)}, \quad W \mapsto \sum_{x \in S} \mathrm{len}(\mathcal{O}_{W,x})[x]$$

is a morphism. The symmetric power $S^{(n)}$ is singular, as for instance the fix-locus of any transposition in G_n has codimension 2. On the other hand by [22] $S^{[n]}$ is smooth and connected of dimension $2n$ and $\rho : S^{[n]} \to S^{(n)}$ is a resolution of singularities. In fact this is a particularly nice resolution: If Y is a Gorenstein variety, i.e. the dualizing sheaf is a line bundle K_Y, a resolution $f : X \to Y$ of singularities is called *crepant* if it preserves the canonical divisor, that is $f^* K_Y = K_X$. It is easy to see

that $\rho : S^{[n]} \to S^{(n)}$ is crepant. In the special case that S is an abelian surface or a K3 surface one can get a better result: A complex manifold X is called *holomorphic symplectic* if there exists an everywhere non-degenerate holomorphic 2-form ϕ on X. If furthermore ϕ is unique up to scalar, X is called irreducible holomorphic symplectic. A Kähler manifold X of real dimension $4n$ is called *hyperkähler* if its holonomy group is $Sp(n)$. Compact complex manifolds are holomorphic symplectic if and only of they admit a hyperkähler metric. In [7] it is shown that for a K3 surface S the Hilbert scheme $S^{[n]}$ is irreducible holomorphic symplectic. There also, for an abelian surface A, the generalized Kummer varieties are constructed from $A^{[n]}$. They form another series of irreducible holomorphic symplectic manifolds. The only other examples of compact hyperkähler manifolds, known not to be diffeomorphic to one in the above two series are the two isolated examples of resolutions of singular moduli spaces of sheaves on K3 and abelian surfaces in [47],[48].

2. Betti number, Euler numbers, elliptic genus

For many questions about the Hilbert schemes $S^{[n]}$ one should look at all n at the same time. The first instance of this are the Betti numbers and Euler numbers, for which we can find generating functions in terms of modular forms. Let $\mathcal{H} := \{\tau \in \mathbf{C} \mid \Im(\tau) > 0\}$. A modular form of weight k on $Sl(2,\mathbf{Z})$ is a function $f : \mathcal{H} \to \mathbf{C}$ s.th.

$$f\left(\frac{a\tau+b}{c\tau+b}\right) = (c\tau+d)^k f(\tau), \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Sl(2,\mathbf{Z}).$$

Furthermore, writing $q = e^{2\pi i\tau}$, we require that, in the Fourier development $f(\tau) = \sum_{n\in\mathbf{Z}} a_n q^n$, all the the negative Fourier coefficients vanish. If also $a_0 = 0$, f is called a cusp form. The most well-known modular form is the discriminant $\Delta(\tau) := q\prod_{n>0}(1-q^n)^{24}$, the unique cusp form of weight 12. The Dirichlet eta function is $\eta = \Delta^{1/24}$.

For a manifold X we denote by $p(X,z) := \sum_i (-1)^i b_i(X) z^i$ the Poincaré polynomial and by $e(X) = p(X,1)$ the Euler number. The Betti numbers and Euler numbers of the $S^{[n]}$ have very nice generating functions [24]:

$$\sum_{n\geq 0} p(S^{[n]},z)t^n = \prod_{k\geq 1}\prod_{i=0}^{4}(1-z^{2k-2+i}t^k)^{(-1)^{i+1}b_i(S)}. \tag{2.1}$$

In particular $\sum_{n\geq 0} e(S^{[n]})q^{n-e(S)/24} = \eta(\tau)^{-e(S)}$.

This was first shown in [19] in the case of the projective plane and of Hirzebruch surfaces using a natural $\mathbf{C}^*$ action. The proof in [24] uses the Weil Conjectures. An important role in this proof as in all subsequent generalizations and refinements is played by the following natural stratification of $S^{[n]}$ and $S^{(n)}$ parametrized by the set $P(n)$ of partitions of n. For a partition $\alpha = (n_1,\ldots,n_r) \in P(n)$, the corresponding locally closed stratum $S^{(n)}_\alpha$ of $S^{(n)}$ consists of the set of zero cycles $n_1[x_1] + \ldots + n_r[x_r]$ with $x_1,\ldots,x_r$ distinct points of S. We put $S^{[n]}_\alpha = \rho^{-1}(S^{(n)}_\alpha)$.

A partition $\alpha = (n_1, \ldots, n_r) \in P(n)$ can also be written as $\alpha = (1^{\alpha_1}, \ldots, n^{\alpha_n})$, where α_i is the number of occurences of i in $(n_1, \ldots, n_r)$. We put $|\alpha| = r = \sum \alpha_i$. Then (2.1) can be reformulated as

$$p(S^{[n]}, z) = \sum_{\alpha \in P(n)} p(S^{(\alpha_1)} \times \ldots \times S^{(\alpha_n)}, z) z^{2(n-|\alpha|)}. \tag{2.2}$$

This result has been refined to Hodge numbers in [30], [11] and this was generalized in [13] to the Douady space of a complex surface. It has been further refined to determine the motive and the Chow groups [14] and the element in the Grothendieck group of varieties of $S^{[n]}$ [28].

Partially motivated by (2.1) and using arguments from physics in [15] a conjectural refinement to the Krichever-Höhn elliptic genus is given. We restrict our attention to the case that $K_X = 0$ when the elliptic genus is a Jacobi form. For a complex vector bundle E on a complex manifold X and a variable t we put

$$\Lambda_t(E) := \bigoplus_{k \geq 0} \Lambda^k(E) t^k, \; S_t(E) := \bigoplus_{k \geq 0} S^k(E) t^k.$$

For the holomorphic Euler characteristic we write $\chi(X, \Lambda_t(E)) := \sum \chi(X, \Lambda^k E) t^k$ and similarly for $S_t(E)$. Then the elliptic genus is defined by

$$\phi(X, q, y) := \chi\left(X, \prod_{m \geq 1} \Lambda_{-y^{-1}q^m} T_X \otimes \Lambda_{-yq^{m-1}} T_X^* \otimes S_{q^m}(T_X \oplus T_X^*)\right).$$

Writing $\phi(S) := \sum_{m \geq 0, l} c(m,l) q^m y^l$, the conjecture is

$$\sum_{N \geq 0} \phi(S^{[n]}) p^N = \prod_{n > 0, m \geq 0, l} \frac{1}{(1 - p^n q^m y^l)^{c(nm,l)}}.$$

3. Infinite dimensional Lie algebras and the cohomology ring

We saw that one gets nice generating functions in n for the Betti numbers of the $S^{[n]}$. Now we shall see that the direct sum of all the cohomologies of the $S^{[n]}$ carries a new structure which governs the ring structures of the Hilbert schemes. We only consider cohomology with rational coefficients and thus write $H^*(X)$ for $H^*(X, \mathbf{Q})$. We write $H := H^*(S)$; for $n > 0$ let $\mathbf{H}_n := H^*(S^{[n]})$. and $\mathbf{H} := \bigoplus_{n \geq 0} \mathbf{H}_n$. We shall see that $\mathbf{H}$ is an irreducible module under a Heisenberg algebra. This was conjectured in [54] and proven in [45],[32]. $\mathbf{H}$ contains a distingished element $\mathbf{1} \in \mathbf{H_0} = \mathbf{Q}$. We denote by $\int_S$ and $\int_{S^{[n]}}$ the evaluation on the fundamental class of S and $S^{[n]}$. Define for $n > 0$ the incidence variety

$$Z_{l,n} := \left\{(Z, x, W) \in S^{[l]} \times S \times S^{[l+n]} \;\middle|\; Z \subset W, \rho(W) - \rho(Z) = n[x]\right\},$$

and use this to define operators

$$p_n : H \to End(\mathbf{H});\ p_n(\alpha)(y) := pr_{3*}(pr_2^*(\alpha) \cup pr_1^*(y) \cap [Z_{l,n}]).$$

Let $p_{-n}(\alpha) := (-1)^n p_{-n}(\alpha)^\dagger$, where † denotes the adjoint with respect to $\int_{S^{[n]}}$, and $p_0(\alpha) := 0$. By [45],[32] the $p_n(\alpha)$ fulfill the commutation relations of a Heisenberg algebra:

$$[p_n(\alpha), p_m(\beta)] = (-1)^{n-1} n \delta_{n,-m} \Big(\int_S \alpha \cdot \beta \Big) id_{\mathbf{H}}, \qquad n, m \in \mathbf{Z},\ \alpha, \beta \in H. \quad (3.1)$$

We can interpret this as follows. Let $H = H_+ \oplus H_-$ be the decomposition into even and odd cohomology. Put $S^*(H) := \bigoplus_{i \geq 0} S^i(H_+) \otimes \bigoplus_{i \geq 0} \Lambda^i(H_-)$. The Fock space associated to H is $F(H) := S^*(H \otimes t\mathbf{Q}[t])$. Using the above theorems one readily shows that there is an isomorphism of graded vector spaces $F(H) \to \mathbf{H}$. With this $\mathbf{H}$ becomes an irreducible module under the Heisenberg-Clifford algebra.

The ring structure of the $H^*(S^{[n]})$ is connected to the Heisenberg algebra action. Given an action of a Heisenberg algebra, a standard construction gives an action of the corresponding Virasoro algebra. The important fact however, proven in [37] is that the Virasoro algebra generators have a geometrical interpretation tying them to the ring structure of the cohomology of the $S^{[n]}$. Let $\delta : S \to S \times S$ be the diagonal embedding, and let $\delta_* : H^*(S) \to H^*(S \times S)$ be the corresponding pushforward. Let $p_\nu p_{n-\nu}\delta(\alpha) : H^*(S) \to End(\mathbf{H})$ be defined as $p_\nu p_{n-\nu}(\beta \times \gamma) := p_\nu(\beta) p_{n-\nu}(\gamma)$ applied to $\delta_*(\alpha) \in H \times H$. For $n \neq 0$ define $L_n(\alpha) := \sum_{\nu \in \mathbf{Z}} p_\nu p_{n-\nu} \delta_*(\alpha)$, and $L_0(\alpha) := \sum_{\nu > 0} p_\nu p_{-\nu} \delta_*(\alpha)$. These operators satisfy the relations of the Virasoro algebra:

$$[L_n(\alpha), L_m(\beta)] = (n-m) L_{n+m}(ab) + \delta_{n,-m} \frac{n^3 - n}{12} \Big(\int_S c_2(S) ab \Big) id_{\mathbf{H}}. \quad (3.2)$$

Let $\partial : \mathbf{H} \to \mathbf{H}$ be the operator which on each $H^*(S^{[n]})$ is the multiplication with $c_1(\mathcal{O}^{[n]})$, where $\mathcal{O}^{[n]} = \pi_*(Z_n(S))$ is the tautological vector bundle associated to the trivial line bundle on S. The tie given in [37] to the ring structure is:

$$[\partial, p_n(\alpha)] = n L_n(a) + \binom{n}{2} p_n(K_S a), \qquad n \geq 0,\ \alpha \in H^*(S). \quad (3.3)$$

In [42], for each $\alpha \in H^*(S)$, classes $\alpha^{[n]} \in H^*(S^{[n]})$ are defined as generatizations of the Chern characters $ch(F^{[n]})$ of tautological bundles, which are studied in [37]. The homogeneous components of the $\alpha^{[n]}$ generate the ring $H^*(S^{[n]})$. [37],[42] relate the multiplication by the $\alpha^{[n]}$ to the higher order commutators with ∂: Let $\alpha^{[\bullet]} : \mathbf{H} \to \mathbf{H}$ be the operator which on every $H^*(S^{[n]})$ is the multiplication with $\alpha^{[n]}$, then

$$[\alpha^{[\bullet]}, p_1(\beta)] = \exp(ad(\partial)) p_1(\alpha\beta), \quad (3.4)$$

where for an operator $A : \mathbf{H} \to \mathbf{H}$, $ad(\partial)A = [\partial, A]$.

(3.2),(3.3),(3.4) determine the cohomology rings of the $S^{[n]}$. In case $K_S = 0$ this is used in [38],[39] to give an elementary description of the cohomology rings $H^*(S^{[n]})$ in terms of the symmetric group, which we will relate below to orbifold cohomology rings.

4. Orbifolds and orbifold cohomology

Let X be a compact complex manifold with an action of a finite group G and assume that for all $1 \neq g \in G$ the fixlocus X^g has codimension ≥ 2. The quotient X/G will usually be singular, but the stack quotient $[X/G]$ is a smooth orbifold. In physics [16],[17] the following orbifold Euler characteristic has been introduced

$$e(X,G) := \sum_{gh=hg \in G} e(X^{g,h}) = \sum_{[g] \subset G} e(X^g/C(g)).$$

Here the first sum runs over all commuting pairs in G and $X^{g,h}$ is the set of common fixpoints; the second sum runs over the conjugacy classes $[g]$ of elements in G and $C(g)$ is the centralizer of g. If $Y \to X/G$ is a crepant resolution, then it was expected that $e(X,G) = e(Y)$. As the conjugacy classes of the symmetric group G_n correspond to the partitions of n, one can see [33] using formula (2.2) that this is true for the resolution of $S^{(n)}$ by $S^{[n]}$, which was an important check for this conjecture.

Orbifold Euler numbers have been refined to orbifold cohomology groups [60]. We again take all cohomology with $\mathbf{Q}$ coefficients. Define a rationally graded $\mathbf{Q}$-vector space

$$H^*_{orb}([X/G]) := \bigoplus_{[g] \subset G} H^*(X^g/C(g)).$$

The grading is defined as follows. Assume for simplicity that all X^g are connected. For $x \in X^g$ let $e^{2\pi i r_1}, \ldots, e^{2\pi i r_k}$ be the eigenvalues of g on $T_{X,x}$. Put $a(g) := \sum r_i \in \mathbf{Q}$ where $r_i \in [0,1)$. This is independent of x. For $\alpha \in H^i(X^g/C(g))$ its degree in the $[g]$-th summand of $H^*_{orb}([X/G])$ is $i + 2a(g)$. If X/G is Gorenstein, then it is easy to see that $a(g) \in \mathbf{Z}_{\geq 0}$. For crepant resolutions $Y \to X/G$, it was conjectured that $H^*_{orb}([X/G]) = H^*(Y)$ as graded vector spaces. In the case of $S^{[n]} \to S^{(n)}$ this can again be verified from formula (2.2). In [6] it has been established for all crepant resolutions $Y \to X/G$.

Recently orbifold cohomology rings, i.e. a ring structure on the orbifold cohomology have been defined as a special case of quantum cohomology of orbifolds [12],[2],[1]. In [50] it is conjectured for an orbifold X with a hyperkähler resolution $Y \to X$, i.e. a crepant resolution such that Y is hyperkähler, that the orbifold cohomology ring of X is isomorphic to $H^*(Y)$. The most relevant case of such a resolution is $S^{[n]} \to S^{(n)}$ when $K_S = 0$. This is precisely the case in which [39] gives an elementary description of the cohomology ring of $S^{[n]}$. In [21],[52] an elementary description of the orbifold cohomology of a quotient $[X/G]$ by a finite group is given. We define $H^*(X,G) := \sum_{g \in G} H^*(X^g)$. This carries a G-action by $h \cdot \alpha_g = (h_* \alpha)_{hgh^{-1}}$, and, for a suitable grading on $H^*(X,G)$, it follows that the G invariant part is just $H^*_{orb}([X/G])$ as a graded vector space. In order to define the ring structure on $H^*_{orb}([X/G])$ one therefore defines a ring structure on $H^*(X,G)$ compatible with the G-action. In [39] the cohomology ring $H^*(S^{[n]})$ is also described as the G_n invariant part of a ring structure on $H^*(S^n, G_n)$ and one checks that the two ring structures on $H^*(S^n, G_n)$ coincide up to an explicit sign change, thus proving the conjecture of [50] for $S^{[n]}$.

If $\pi : Y \to X/G$ is only a crepant resolution but not hyperkähler, then usually $H^*_{orb}([X/G])$ and $H^*(Y)$ are not isomorphic as rings. However in [51] a precise conjecture is made relating the two: One has to correct $H^*_{orb}([X/G])$ by Gromov-Witten invariants coming from classes of rational curves Y contracted by π. In the case of the Hilbert scheme these curve classes are the multiples of a unique class. The conjecture was verified for $S^{[2]}$.

5. Moduli of vector bundles

We denote by $M_S^H(r, c_1, c_2)$ the moduli space of Gieseker H-semistable coherent sheaves of rank r on S with Chern classes c_1, c_2. Here a sheaf $\mathcal{F}$ of rank $r > 0$ on S is called semistable, if $\chi(\mathcal{G} \otimes H^n)/r' \leq \chi(\mathcal{F} \otimes H^n)/r$ for all sufficiently large n and for all subsheaves $\mathcal{G} \subset \mathcal{F}$ of positive rank r'. As $M_S^H(1, 0, c_2) \simeq Pic^0(S) \times S^{[c_2]}$, the Hilbert scheme of points is a special case. We will often restrict our attention to the case of $r = 2$ and write $M_S^H(c_1, c_2)$. The Hilbert schemes of points are related in several ways to the $M_S^H(c_1, c_2)$. The most basic tie is the Serre correspondence which says that under mild assumptions rank two vector bundles on S can be constructed as extensions of ideal sheaves of finite subschemes by line bundles. Related to this is the dependence of the $M_S^H(c_1, c_2)$ on the ample divisor H via a system of walls and chambers. This has been studied by a number of authors (e.g. [49],[23],[18]). Assume for simplicity that S is simply connected. A class $\xi \in H^2(S, \mathbf{Z})$ defines a wall of type (c_1, c_2) if $\xi + c_1 \in 2H^2(S, \mathbf{Z})$ and $c_1^2 - 4c_2 \leq \xi^2 < 0$. The corresponding wall is $W^\xi = \{\alpha \in H^2(S, \mathbf{R}) \mid \alpha \cdot \xi = 0\}$. The connected components of the complement of the walls in $H^2(X, \mathbf{R})$ are called the chambers of type (c_1, c_2). If a sheaf $\mathcal{E} \in M_S^H(c_1, c_2)$ is unstable with respect to L, then there is a wall W^ξ with $H\xi < 0 < L\xi$ and an extension

$$0 \to \mathcal{I}_Z \otimes A \to \mathcal{E} \to \mathcal{I}_W \times B \to 0,$$

where $A, B \in Pic(S)$ with $A - B = \xi$ and $\mathcal{I}_Z$, $\mathcal{I}_W$ are the ideal sheaves of zero dimensional schemes on S. It follows that $M_S^H(c_1, c_2)$ depends only on the chamber of H and the set theoretic change under wallcrossing is given in terms of Hilbert schemes of points on S. In the case e.g. of rational surfaces and K3-surfaces, the change can be described as an explicit sequence of blow ups along $\mathbf{P}_k$ bundles over products $S^{[n]} \times S^{[m]}$ followed by blow downs in another direction [23],[18]. The change of the Betti and Hodge numbers under wallcrossing can be explicitly determined and this can be used e.g. to determine the Hodge numbers of $M_S^H(c_1, c_2)$ for rational surfaces. For suitable choices of H one can find the generating functions in terms of modular forms and Jacobi forms [27].

The appearance of modular forms is in accord with the S-duality conjectures [54] from theoretical physics, which predict that under suitable assumptions the generating functions for the the Euler numbers of moduli spaces of sheaves on surfaces should be given by modular forms. One of the motivating examples for this conjecture is the case that S is a K3-surface. In this case the conjecture is that, if $M_S^H(c_1, c_2)$ is smooth, then it has the same Betti numbers as the Hilbert scheme of points on S of the same dimension. Assuming this, the formula (2.1) for the Hilbert

schemes of points implies that the generating function for the Euler numbers is a modular form. If c_1 is primitive this was shown in [29]. The result was shown in general for $M_S^H(r, c_1, c_2)$ with $r > 0$ in [57],[59], by relating the Hilbert scheme and the moduli space via birational correspondences and deformations. One concludes that $M_S^H(r, c_1, c_2)$ has the same Betti numbers as the Hilbert scheme of points of the same dimension, as both spaces are holomorphic symplectic [44] and birational manifolds with trivial canonical class have the same Betti numbers [5]. Similar results are shown in [58] for abelian surfaces. Other motivating examples for the S-duality conjecture were the case of $\mathbf{P}_2$ [56] and the blowup formula relating the generating function for the Euler numbers of the moduli spaces of rank 2 sheaves on a surface S to that on the blowup of S in a point, which has since been established ([40],[41], see also [27]).

The moduli spaces $M_S^H(c_1, c_2)$ can be used to compute the Donaldson invariants of S. In case $p_g = 0$ these depend on a metric, corresponding to the dependence of $M_S^H(c_1, c_2)$ on H. For rational surfaces one can use the above description of the wallcrossing for the $M_S^H(c_1, c_2)$ to determine the change of the Donaldson invariants in terms of Chern numbers of generalizations of the tautological sheaves $L^{[n]}$ on products $S^{[n]} \times S^{[m]}$ of Hilbert schemes of points [18],[23]. The leading terms of these expressions can be explicitly evaluated. The wallcrossing of Donaldson invariants has also been studied in gauge theory (e.g.[35],[36]). There a conjecture about the structure of the wallcrossing formulas is made. Assuming this conjecture one can determine the generating functions for the wallcrossing in terms of modular forms [25],[31].

6. Enumerative geometry of curves

Now we want to see some striking relations between the Hilbert schemes $S^{[n]}$ and the enumerative geometry of curves on S. First let S be a K3 surface and L a primitive line bundle on S. Then $L^2 = 2g - 2$, where the linear system $|L|$ has dimension g and a smooth curve in $|L|$ has geometric genus g. As a node imposes one linear condition, one expects a finite number of rational curves (i.e. curves of geometric genus 0) in $|L|$. Partially based on arguments from physics, a formula is given in [55] for the number of rational curves in $|L|$ and in [8] this made mathematically precise. Writing n_g for the number of rational curves in $|L|$ with $L^2 = 2g - 2$ (counted with suitable multiplicities), the formula is

$$\sum_{g \geq 0} n_g q^g = \frac{q}{\Delta}, \tag{6.1}$$

where Δ is again the discriminant. By (2.1) this implies the surprizing fact that n_g is just the Euler number of $S^{[g]}$. In fact the argument relates the number of curves to $S^{[g]}$: Let $\mathcal{C} \to |L|$ be the universal curve and let $\mathcal{J} \to |L|$ be the corresponding relative compactified Jacobian, whose fibre over the point corresponding to a curve C is the compactified Jacobian $J(C)$ [3]. One can show that $e(J(C)) = 0$ unless $g(C) = 0$. It follows that $e(\mathcal{J})$ is the sum over the $e(J(C))$ for $C \in |L|$ with $g(C) = 0$. It is not difficult to show that $S^{[g]}$ and $\mathcal{J}$ are birational. $\mathcal{J}$ is also smooth

and hyperkähler as a moduli space of sheaves on a K3 surface [44]. As already used in the section on vector bundles, birational manifolds with trivial canonical bundle have the same Betti numbers [5]. Thus $\mathcal{J}$ and $S^{[g]}$ have the same Euler numbers. This shows (6.1), where the multiplicity of a rational curve C is $e(J(C))$. By [20] this multiplicity is the multiplicity of the corresponding moduli space of stable maps, in particular it is always positive. In [26] a conjectural generalization of (6.1) to arbitrary surfaces S is given.

Conjecture 6.1

1. *For all $\delta \geq 0$, there exists a universal polynomial $T_\delta(x,y,z,w)$, such that for all projective surfaces S and all sufficiently ample line bundles L on S the number of δ-nodal curves in a general δ-dimensional linear subspace of $|L|$ is $T_\delta(\chi(L),\chi(\mathcal{O}_S),LK_S,K_S^2)$.*
2. *There are universal power series $B_1, B_2 \in \mathbf{Z}[[q]]$ whose coefficients can be explicitely determined, such that*

$$\sum_{\delta\geq 0} T_\delta(\chi(L),\chi(\mathcal{O}_S),LK_S,K_S^2)(DG_2)^\delta = \frac{(DG_2/q)^{\chi(L)}B_1^{LK_S}B_2^{K_S^2}}{(\Delta D^2 G_2/q^2)^{\chi(\mathcal{O}_S)/2}}.$$

Here $D = q\frac{d}{dq}$ and $G_2 = -\frac{1}{24}\frac{D\Delta}{\Delta}$.

The expectation that universal polynomials should exist is implicit in [53],[34] where the T_δ are determined for $\delta \leq 8$. In [26] also another tie of the conjecture to the Hilbert scheme of points is given: conjecturally the numbers $T_\delta(\chi(L),\chi(\mathcal{O}_S), LK_S, K_S^2)$ are suitable intersection numbers on the Hilbert scheme $S^{[3\delta]}$ of 3δ points of S. If S is a K3 surface or an abelian surface, then the conjecture predicts that the generating function can be written in terms of modular forms. In this case a modified version of Conjecture 6.1 was proven for primitive line bundles in [9] and [10], replacing the numbers of δ-nodal curves with the corresponding modified Gromov-Witten invariants. In [43] a proof of the conjecture is published.

References

[1] D. Abramovich, T. Graber, A. Vistoli, Algebraic orbifold quantum products, preprint math.AG/0112004.

[2] D. Abramovich, A. Vistoli, Compactifying the space of stable maps, *J. Amer. Math. Soc.*, 15 (2002), 27–75.

[3] A. Altman, S. Kleiman, Compactifying the Picard scheme. *Adv. in Math.*, 35 (1980), 50–112.

[4] A. Altman, S. Kleiman, Compactifying the Picard scheme II, *Amer. J. Math.*, 101 (1979), 10–41.

[5] V. Batyrev, Birational Calabi-Yau n-folds have equal Betti numbers. *New trends in algebraic geometry*, 1–11, 1999.

[6] V. Batyrev, L. Borisov, Mirror duality and string-theoretic Hodge numbers, *Invent. Math.*, 126 (1996), 183–203.

[7] A. Beauville, Variétés Kähleriennes dont la première classe de Chern est nulle, *J. Differential Geom.*, 18 (1983), 755–782.
[8] A. Beauville, Counting rational curves on $K3$ surfaces, *Duke Math. J.*, 97 (1999), 99–108.
[9] J. Bryan, N. C. Leung, The enumerative geometry of $K3$ surfaces and modular forms. *J. Amer. Math. Soc.*, 13 (2000), 371–410.
[10] J. Bryan, N. C. Leung, Generating functions for the number of curves on abelian surfaces, *Duke Math. J.*, 99 (1999), 311–328.
[11] J. Cheah, On the cohomology of Hilbert schemes of points, *J. Algebraic Geom.*, 5 (1996), 479–511.
[12] W. Chen, Y. Ruan, A New Cohomology Theory for Orbifold, preprint math.AG/0004129.
[13] M. A. de Cataldo, L. Migliorini, The Douady space of a complex surface, *Adv. Math.*, 151 (2000), 283–312.
[14] M. A. de Cataldo, L. Migliorini, The Chow groups and the motive of the Hilbert scheme of points on a surface, preprint math.AG/0005249.
[15] R. Dijkgraaf, G. Moore, E. Verlinde, H. Verlinde, Elliptic genera of symmetric products and second quantized strings, *Comm. Math. Phys.*, 185 (1997), 197–209.
[16] L. Dixon, J. Harvey, C. Vafa, E. Witten, Strings on orbifolds, *Nuclear Phys. B*, 261 (1985), 678–686.
[17] L. Dixon, J. Harvey, C. Vafa, E. Witten, Strings on orbifolds II, *Nuclear Phys. B* 274 (1986), 285–314.
[18] G. Ellingsrud, L. Göttsche, Variation of moduli spaces and Donaldson invariants under change of polarization, *J. Reine Angew. Math.*, 467 (1995), 1–49.
[19] G. Ellingsrud, S. A. Strømme, On the homology of the Hilbert scheme of points in the plane, *Invent. Math.*, 87 (1987), 343–352.
[20] B. Fantechi, L. Göttsche, D. van Straten, Euler number of the compactified Jacobian and multiplicity of rational curves *J. Algebraic Geom.*, 8 (1999), 115–133.
[21] B. Fantechi, L. Göttsche, Orbifold cohomology for global quotients, preprint math.AG/0104207.
[22] J. Fogarty, Algebraic families on an algebraic surface, *Amer. J. Math.*, 90 (1968), 511–521.
[23] R. Friedman, Z. Qin, Flips of moduli spaces and transition formulas for Donaldson polynomial invariants of rational surfaces, *Comm. Anal. Geom.*, 3 (1995), 11–83.
[24] L. Göttsche, The Betti numbers of the Hilbert scheme of points on a smooth projective surface, *Math. Ann.* 286 (1990), 193–207.
[25] L. Göttsche, Modular forms and Donaldson invariants for 4-manifolds with $b_+ = 1$, *J. Amer. Math. Soc.*, 9 (1996), 827–843.
[26] L. Göttsche, A conjectural generating function for numbers of curves on surfaces, *Comm. Math. Phys.*, 196 (1998), 523–533.
[27] L. Göttsche, Theta functions and Hodge numbers of moduli spaces of sheaves on rational surfaces, *Comm. Math. Phys.*, 206 (1999), 105–136.

[28] L. Göttsche, On the motive of the Hilbert scheme of points on a surface, *Math. Res. Lett.*, 8 (2001), 613–627.
[29] L. Göttsche, D. Huybrechts, Hodge numbers of moduli spaces of stable bundles on $K3$ surfaces, *Internat. J. Math.*, 7 (1996), 359–372.
[30] L. Göttsche, W. Soergel, Perverse sheaves and the cohomology of Hilbert schemes of smooth algebraic surfaces, *Math. Ann.*, 296 (1993), 235–245.
[31] L. Göttsche, D. Zagier, Jacobi forms and the structure of Donaldson invariants for 4-manifolds with $b_+ = 1$. *Selecta Math. (N.S.)*, 4 (1998), 69–115.
[32] I. Grojnowski, Instantons and affine algebras. I. The Hilbert scheme and vertex operators, *Math. Res. Lett.*, 3 (1996), 275–291.
[33] F. Hirzebruch, T. Höfer, On the Euler number of an orbifold, *Math. Ann.*, 286 (1990), 255–260.
[34] S. Kleiman, R. Piene, Enumerating singular curves on surfaces, *Algebraic geometry: Hirzebruch 70*, 209–238, Contemp. Math., 241, 1999.
[35] D. Kotschick, SO(3)-invariants for 4-manifolds with $b_2^+ = 1$, *Proc. London Math. Soc.*, 63 (1991), 426–448.
[36] D. Kotschick, D.; J. Morgan, SO(3)-invariants for 4-manifolds with $b_2^+ = 1$, II. *J. Differential Geom. 39*, (1994), 433–456.
[37] M. Lehn, Chern classes of tautological sheaves on Hilbert schemes of points on surfaces, *Invent. Math.*, 136 (1999), 157–207.
[38] M. Lehn, C. Sorger, Symmetric groups and the cup product on the cohomology of Hilbert schemes, *Duke Math. J.*, 110 (2001), 345–357.
[39] M. Lehn, C. Sorger, The cup product of the Hilbert scheme for K3 surfaces, preprint math.AG/0012166.
[40] W. Li, Z. Qin, Li, On blowup formulae for the S-duality conjecture of Vafa and Witten, *Invent. Math.*, 136 (1999), 451–482.
[41] W. Li, Z. Qin, On blowup formulae for the S-duality conjecture of Vafa and Witten. II, *Math. Res. Lett.*, 5 (1998), 439–453.
[42] W. Li, Z. Qin, W. Wang, Vertex algebras and the cohomology ring structure of Hilbert schemes of points on surfaces, preprint math.AG/0009132.
[43] A. K. Liu, Family blowup formula, admissible graphs and the enumeration of singular curves I, *J. Differential Geom.*, 56 (2000), 381–579.
[44] S. Mukai, Symplectic structure of the moduli space of sheaves on an abelian or $K3$ surface, *Invent. Math.*, 77 (1984), 101–116.
[45] H. Nakajima, Heisenberg algebra and Hilbert schemes of points on projective surfaces, *Ann. of Math.*, 145 (1997), 379–388.
[46] H. Nakajima, *Lectures on Hilbert schemes of points on surfaces*, University Lecture Series, 18. American Mathematical Society, 1999.
[47] K. O'Grady, Desingularized moduli spaces of sheaves on a $K3$, *J. Reine Angew. Math.*, 512 (1999), 49–117.
[48] K. O'Grady, A new six dimensional irreducible symplectic variety, preprint math.AG/0010187.
[49] Z. Qin, Equivalence classes of polarizations and moduli spaces of sheaves, *J. Differential Geom.*, 37 (1993), 397–415.
[50] Y. Ruan, Stringy Geometry and Topology of Orbifolds, preprint

math.AG/0011149.
[51] Y. Ruan, Cohomology ring of crepant resolutions of orbifolds, preprint math.AG/0108195.
[52] B. Uribe, Orbifold Cohomology of the Symmetric Product, preprint math.AT/0109125.
[53] I. Vainsencher, Enumeration of n-fold tangent hyperplanes to a surface, *J. Algebraic Geom.*, 4 (1995), 503–526.
[54] C. Vafa, E. Witten, A strong coupling test of S-duality, *Nuclear Phys. B*, 431 (1994), 3–77.
[55] S.-T, Yau, E. Zaslow, BPS states, string duality, and nodal curves on $K3$. *Nuclear Phys. B*, 471 (1996), 503–512.
[56] K. Yoshioka, The Betti numbers of the moduli space of stable sheaves of rank 2 on $\mathbf{P}^2$, *J. Reine Angew. Math.*, 453 (1994), 193–220.
[57] K. Yoshioka, Some examples of Mukai's reflections on $K3$ surfaces, *J. Reine Angew. Math.*, 515 (1999), 97–123.
[58] K. Yoshioka, Moduli spaces of stable sheaves on abelian surfaces, *Math. Ann.*, 321 (2001), 817–884.
[59] K. Yoshioka, Irreducibility of moduli spaces of vector bundles on K3 surfaces, preprint math.AG/9907001.
[60] E. Zaslow, Topological orbifold models and quantum cohomology rings, *Comm. Math. Phys.*, 156 (1993), 301–331.

ICM 2002 · Vol. II · 495–502

Vector Bundles on a K3 Surface*

Shigeru Mukai†

Abstract

A K3 surface is a quaternionic analogue of an elliptic curve from a view point of moduli of vector bundles. We can prove the algebraicity of certain Hodge cycles and a rigidity of curve of genus eleven and gives two kind of descriptions of Fano threefolds as applications. In the final section we discuss a simplified construction of moduli spaces.

2000 Mathematics Subject Classification: 14J10, 14J28, 14J60.

1. Introduction

A locally free sheaf E of $\mathcal{O}_X$-modules is called *a vector bundle* on an algebraic variety X. As a natural generalization of line bundles vector bundles have two important roles in algebraic geometry. One is the linear system. If E is generated by its global sections $H^0(X,E)$, then it gives rise to a morphism to a Grassmann variety, which we denote by $\Phi_E : X \longrightarrow G(H^0(E),r)$, where r is the rank of E. This morphism is related with the classical linear system by the following diagram:

$$\begin{array}{ccccc} & X & \stackrel{\Phi_E}{\longrightarrow} & G(H^0(E),r) & \\ \Phi_L & \downarrow & & \downarrow & \text{Plücker} \\ & \mathbf{P}^*H^0(L) & \cdots\to & \mathbf{P}^*(\textstyle\bigwedge^r H^0(E)), & \end{array} \tag{1}$$

where L is the determinant line bundle of E and Φ_L is the morphism associated to it.

The other role is the moduli. The moduli space of line bundles relates a (smooth complete) algebraic curve with an abelian variety called the *Jacobian variety*, which is crucial in the classical theory of algebraic functions in one variable. The moduli of vector bundles also gives connections among different types varieties, and often yields new varieties that are difficult to describe by other means.

*Supported in part by the JSPS Grant-in-Aid for Scientific Research (A) (2) 10304001.

†Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, Japan. E-mail: mukai@kurims.kyoto-u.ac.jp

In higher rank case it is natural to consider the moduli problem of E under the restriction that $\det E$ is unchanged. In view of the above diagram, vector bundles and their moduli reflect the geometry of the morphism $X \longrightarrow \mathbf{P}^* H^0(L)$ via Grassmannians and Plücker relations. In this article we consider the case where X is a K3 surface, which is one of two 2-dimensional analogues of an elliptic curve and seems an ideal place to see such reflection.

2. Curves of genus one

The moduli space of line bundles on an algebraic variety is called the *Picard variety*. The Picard variety $\operatorname{Pic} C$ of an algebraic curve C is decomposed into the disjoint union $\coprod_{d \in \mathbf{Z}} \operatorname{Pic}_d C$ by the degree d of line bundles. Here we consider the case of genus 1. All components $\operatorname{Pic}_d C$ are isomorphic to C if the ground field is algebraically closed. [1] But this is no more true otherwise. For example the Jacobian $\operatorname{Pic}_0 C$ has always a rational point but C itself does not. [2] We give other examples:

Example 1 Let C_4 be an intersection of two quadrics $q_1(x) = q_2(x) = 0$ in the projective space $\mathbf{P}^3$ and P the pencil of defining quadrics. Then the Picard variety $\operatorname{Pic}_2 C_4$ is the double cover of $P \simeq \mathbf{P}^1$ and the branch locus consists of 4 singular quadrics in P. Precisely speaking, its equation is given by $\tau^2 = \operatorname{disc}(\lambda_1 q_1 + \lambda_2 q_2)$.

Let $G(2,5) \subset \mathbf{P}^9$ be the 6-dimensional Grassmann variety embedded into $\mathbf{P}^9$ by the Plücker coordinate. Its projective dual is the dual Grassmannian $G(5,2) \subset \hat{\mathbf{P}}^9$, where $G(2,5)$ parameterizes 2-dimensional subspaces and $G(5,2)$ quotient spaces.

Example 2 A transversal linear section $C = G(2,5) \cap H_1 \cap \cdots \cap H_5$ is a curve genus 1 and of degree 5. Its Picard variety $\operatorname{Pic}_2 C$ is isomorphic to the dual linear section $\hat{C} = G(5,2) \cap \langle H_1, \ldots, H_5 \rangle$, the intersection with the linear subspace spanned by 5 points $H_1, \ldots, H_5 \in \hat{\mathbf{P}}^9$.

3. Moduli K3 surfaces

A compact complex 2-dimensional manifold S is *a K3 surface* if the canonical bundle is trivial and the irregularity vanishes, that is, $K_S = H^1(\mathcal{O}_S) = 0$. A smooth quartic surface $S_4 \subset \mathbf{P}^3$ is the most familiar example. Let us first look at the 2-dimensional generalization of Example 1:

Example 3 Let S_8 be an intersection of three general quadrics in $\mathbf{P}^5$ and N the net of defining quadrics. Then the moduli space $M_S(2, \mathcal{O}_S(1), 2)$ is a double cover of $N \simeq \mathbf{P}^2$ and the branch locus, which is of degree 6, consists of singular quadrics in N.

Here $M_S(r, L, s)$, L being a line bundle, is the moduli space of stable sheaves E on a K3 surface S with rank r, $\det E \simeq L$ and $\chi(E) = r + s$. Surprisingly two

[1] More precisely, this holds true if C has a rational point.

[2] Two components $\operatorname{Pic}_0 C$ and $\operatorname{Pic}_{g-1} C$ deserve the name *Jacobian*. They coincide in our case $g = 1$.

surfaces $S_8 = (2) \cap (2) \cap (2) \subset \mathbf{P}^5$ and $M_S(2, \mathcal{O}_S(1), 2) \xrightarrow{2:1} \mathbf{P}^2$ in this example are both K3 surfaces. This is not an accident. In respect of moduli space, vector bundles a K3 surface look like Picard varieties in the preceding section.

Theorem 1 ([10],[11]) *The moduli space $M_S(r, L, s)$ is smooth of dimension $(L^2) - 2rs + 2$. $M_S(r, L, s)$ is again a K3 surface if it is compact and and of dimension 2.*

A K3 surface S and a moduli K3 surface appearing as $M_S(r, L, s)$ are not isomorphic in general [3] but their polarized Hodge structures, or periods, are isomorphic to each other over $\mathbf{Q}$ ([11]). The moduli is not always fine but there always exists a universal $\mathbf{P}^{r-1}$-bundle over the product $S \times M_S(r, L, s)$. Let $\mathcal{A}$ be the associated sheaf of Azumaya algebras, which is of rank r^2 and locally isomorphic to the matrix algebra $Mat_r(\mathcal{O}_{S \times M})$. $\mathcal{A}$ is isomorphic to $\mathcal{E}nd\,\mathcal{E}$ if a universal family $\mathcal{E}$ exists. The Hodge isometry between $H^2(S, \mathbf{Q})$ and $H^2(M_S(r, L, s), \mathbf{Q})$ is given by $c_2(\mathcal{A})/2r \in H^4(S \times M, \mathbf{Q}) \simeq H^2(S, \mathbf{Q})^\vee \otimes H^2(M, \mathbf{Q})$.

Example 2 has also a K3 analogue. Let $\Sigma_{12} \subset \mathbf{P}^{15}$ be a 10-dimensional spinor variety $SO(10)/U(5)$, that is, the orbit of a highest weight vector in the projectivization of the 16-dimensional spinor representation. The anti-canonical class is 8 times the hyperplane section and a transversal linear section $S = \Sigma_{12} \cap H_1 \cap \cdots \cap H_8$ is a K3 surface (of degree 12). As is similar to $G(2, 5)$ the projective dual $\hat{\Sigma}_{12} \subset \hat{\mathbf{P}}^{15}$ of Σ_{12} is again a 10-dimensional spinor variety.

Example 4 The moduli space $M_S(2, \mathcal{O}_S(1), 3)$ is isomorphic to the dual linear section $\hat{S} = \hat{\Sigma}_{12} \cap \langle H_1, \ldots, H_8 \rangle$.

In this case, moduli is fine and the relation between S and the moduli K3 are deeper. The universal vector bundle $\mathcal{E}$ on the product gives an equivalence between the derived categories $\mathbf{D}(S)$ and $\mathbf{D}(\hat{S})$ of coherent sheaves, the duality $\hat{\hat{S}} \simeq S$ holds (cf. [17]) and moreover the Hilbert schemes $\mathrm{Hilb}^2\, S$ and $\mathrm{Hilb}^2\, \hat{S}$ are isomorphic to each other.

Remark (1) Theorem 1 is generalized to the non-compact case by Abe [1].
(2) If a universal family $\mathcal{E}$ exists, the derived functor with kernel $\mathcal{E}$ gives an equivalence of derived categories of coherent sheaves on S and the moduli K3 (Bridgeland [4]). In even non-fine case the derived category $\mathbf{D}(S)$ is equivalent to that of the moduli K3 M twisted by a certain element $\alpha \in H^2(M, \mathbf{Z}/r\mathbf{Z})$ (Căldăraru [5]).

4. Shafarevich conjecture

Let S and T be algebraic K3 surfaces and f a Hodge isometry between $H^2(S, \mathbf{Z})$ and $H^2(T, \mathbf{Z})$. Then the associated cycle $Z_f \in H^4(S \times T, \mathbf{Z}) \simeq H^2(S, \mathbf{Z})^\vee \otimes H^2(T, \mathbf{Q})$ on the product $S \times T$ is a Hodge cycle. This is algebraic by virtue of the Torelli type theorem of Shafarevich and Piatetskij-Shapiro. Shafarevich conjectured in [23] a generalization to Hodge isometries over $\mathbf{Q}$. Our moduli theory is able to answer this affirmatively.

[3] We take the complex number field $\mathbf{C}$ as ground field except for sections 2 and 7.

Theorem 2 *Let $f : H^2(S, \mathbf{Q}) \longrightarrow H^2(T, \mathbf{Q})$ be a Hodge isometry. Then the associated (Hodge) cycle $Z_f \in H^4(S \times T, \mathbf{Q})$ is algebraic.*

In [11], we already proved this partially using Theorem 1 (cf. [21] also). What we need further is the moduli space of projective bundles. Let $P \longrightarrow S$ be a $\mathbf{P}^{r-1}$-bundle over S. The cohomology class $[P] \in H^1(S, PGL(r, \mathcal{O}_S))$ and the natural exact sequence (in the classical topology)

$$0 \longrightarrow \mathbf{Z}/r\mathbf{Z} \longrightarrow SL(r, \mathcal{O}_S) \longrightarrow PGL(r, \mathcal{O}_S) \longrightarrow 1$$

define an element of $H^2(S, \mathbf{Z}/r\mathbf{Z})$, which we denote by $w(P)$.

Fix $\alpha \in H^2(S, \mathbf{Z})$ and r, we consider the moduli of $\mathbf{P}^{r-1}$-bundles P over S with $w(P) = \alpha \bmod r$ which are stable in a certain sense. If the self intersection number (α^2) is divisible by $2r$, then the moduli space contains a 2-dimensional component, which we denote by $M_S(\alpha/r)$. The following, a honest generalization of computations in [11], is the key of our proof:

Proposition 1 *Assume that (α^2) is divisible by $2r^2$. Then $H^2(M_S(\alpha/r), \mathbf{Z})$ is isomorphic to $L_0 + \mathbf{Z}\alpha/r \subset H^2(S, \mathbf{Q})$ as polarized Hodge structure, where L_0 is the submodule of $H^2(S, \mathbf{Z})$ consisting of β such that the intersection number $(\beta.\alpha)$ is divisible by r.*

For example let S_2 be a double cover of $\mathbf{P}^2$ with branch sextic. If $\alpha \in H^2(S, \mathbf{Z})$ satisfies $(\alpha.h) \equiv 1 \bmod 2$ and $(\alpha^2) \equiv 0 \bmod 4$, then $M_S(\alpha/2)$ is a K3 surface of degree 8. This is the inverse correspondence of Example 1 (cf. [26], [20]). Details will be published elsewhere.

5. Non-Abelian Brill-Noether locus

Let C be a smooth complete algebraic curve. As a set a Brill-Noether locus of C is a stratum of the Picard variety $\operatorname{Pic} C$ defined by $h^0(L)$, the number of global sections of a line bundle L. The standard notation is

$$W_d^r = \{[L] \,|\, h^0(L) \geq r + 1\} \subset \operatorname{Pic}_d C,$$

for which we refer [2]. Non-Abelian analogues are defined in the moduli space $\mathcal{U}_C(2)$ of stable 2-bundles on C similarly. The *non-Abelian Brill-Noether locus of type III* is

$$\mathcal{SU}_C(2, K : n) = \{F \,|\, \det F \simeq \mathcal{O}_C(K_C),\ h^0(F) \geq n\} \subset \mathcal{U}_C(2)$$

for a non-positive integer n, and *type II* is

$$\mathcal{SU}_C(2, K : nG) = \{F \,|\, \det F \simeq \det G \otimes \mathcal{O}_C(K_C),\ \dim \operatorname{Hom}(G, F) \geq n\} \subset \mathcal{U}_C(2)$$

for a vector bundle G of rank 2 and $n \equiv \deg G \bmod 2$. By virtue of the (Serre) duality, these have very special determinantal descriptions. We give them scheme structures using these descriptions ([16]).

Assume that C lies on a K3 surface S. If E belongs to $M_S(r, L, s)$, then the restriction $E|_C$ is of canonical determinant and we have $h^0(E|_C) \geq h^0(E) \geq \chi(E) = r + s$. So $E|_C$ belongs to $\mathcal{SU}_C(2, K : r + s)$ if it is stable. This is one motivation of the above definition. The case of genus 11, the gap value of genera where Fano 3-folds of the next section do not exist, is the molst interesting.

Theorem 3 ([15]) *If C is a general curve of genus* 11, *then the Brill-Noether locus $T = \mathcal{SU}_C(2, K : 7)$ of type III is a K3 surface and the restriction L of the determinant line bundle is of degree* 20.

There exists a universal family $\mathcal{E}$ on $C \times T$. We moreover have the following:

- the restriction $\mathcal{E}|_{x \times T}$ is is stable and belongs to $M_T(2, L, 5)$, for every $x \in C$, and
- the classification morphism $C \longrightarrow \hat{T} = M_T(2, L, 5)$ is an embedding.

This embedding is a non-Abelian analogue of the Albanese morphism $X \longrightarrow \mathrm{Pic}_0(\mathrm{Pic}_0 X)$ and we have the following:

Corollary *A general curve of genus eleven has a unique embedding to a K3 surface.*

In [9], we studied the forgetful map φ_g from the moduli space $\mathcal{P}_g$ of pairs of a curve C of genus g and a K3 surface S with $C \subset S$ to the moduli space $\mathcal{M}_g$ of curves of genus g and the generically finiteness of φ_{11}. The above correspondence $C \mapsto \hat{T}$ gives the inverse rational map of φ_{11}. We recall the fact that φ_{10} is not dominant in spite of the inequality $\dim \mathcal{P}_{10} = 29 > \dim \mathcal{M}_{10} = 27$ ([12]).

6. Fano 3-folds

A smooth 3-dimensional projective variety is called a *Fano 3-fold* if the anticanonical class $-K_X$ is ample. In this section we assume that the Picard group $\mathrm{Pic}\, X$ is generated by $-K_X$. The self intersection number $(-K_X)^3 = 2g - 2$ is always even and the integer $g \geq 2$ is called the *genus*, by which the Fano 3-folds are classified into 10 deformation types. The values of g is equal to $2, \ldots, 10$ and 12. A Fano 3-folds of genus $g \leq 5$ is a complete intersection of hypersurfaces in a suitable weighted projective space.

By Shokurov [25], the anticanonical linear system $|-K_X|$ always contain a smooth member S, which is a K3 surface. In [13] we classified Fano 3-folds X of Picard number one using *rigid bundles*, that is, $E \in M_S(r, L, s)$ with $(L^2) - 2rs = -2$. For example X is isomorphic to a linear section of the 10-dimensional spinor variety, that is,

$$X \simeq \Sigma_{12} \cap H_1 \cap \cdots \cap H_7, \tag{2}$$

in the case of genus 7 and a linear section

$$X \simeq \Sigma_{16} \cap H_1 \cap H_2 \cap H_3, \tag{3}$$

of the 6-dimensional symplectic, or Lagrangian, Grassmann variety $\Sigma_{16} = SP(6)/U(3) \subset \mathbf{P}^{13}$ in the case of genus 9. The non-Abelian Brill-Noether loci shed new light on this classification.

Theorem 4 *A Fano 3-fold X of genus 7 is isomorphic to the Brill–Noether locus $\mathcal{SU}_C(2, K:5)$ of Type III for a smooth curve C of genus 7.*

This description is dual to the description (2) in the following sense. First two ambient spaces of X, the moduli $\mathcal{U}_C(2)$ and the Grassmannian $G(5,10) \supset \Sigma_{12}$ are of the same dimension. Secondly let N_1 and N_2 be the normal bundles of X in these ambient spaces. Then we have $N_1 \simeq N_2^\vee \otimes \mathcal{O}_X(-K_X)$, that is, two normal bundles are twisted dual to each other.

Theorem 5 *A Fano 3-fold of genus 9 is isomorphic to the Brill–Noether locus $\mathcal{SU}_C(2, K:3G)$ of Type II for a nonsingular plane quartic curve C and G a rank 2 stable vector bundle over C of odd degree.*

This descriptions is also dual to (3) in the above sense: The moduli $\mathcal{U}_C(2)$ and the Grassmannian $G(3,6) \supset \Sigma_{16}$ are of the same dimension and the two normal bundles of X are twisted dual to each other. Each Fano 3-fold of genus 8, 12 and conjecturally 10 has also such a pair of descriptions.

7. Elementary construction

The four examples in sections 1 and 2 are very simple and invite us to a simplification of moduli construction. Let C_4 be as in Example 1 and $\mathcal{M}at_2$ the affine space associated to the 16-dimensional vector space $\bigoplus_{i=0}^3 (\mathbf{C}^2 \otimes \mathbf{C}^2)x_i$, where (x_i) is the homogeneous coordinate of $\mathbf{P}^3$. Let $\mathcal{M}at_{2,1}$ be the closed subscheme defined by the condition that

$$A(x) = \sum_{i=0}^3 A_i x_i \in \mathcal{M}at_2 \text{ is of rank} \leq 1 \text{ everywhere on } C_4$$

and R its coordinat ring. Then the Picard variety $\mathrm{Pic}_2 C_4$ is the projective spectrum $\mathrm{Proj}\, R^{SL(2)\times SL(2)}$ of the invariant ring by construction. (See [18] for details.) The above condition is equivalent to that $\det A(x)$ is a linear combination of $q_1(x)$ and $q_2(x)$. The invariant ring is generated by three elements by Theorem 2.9.A of Weyl [28]. Two of them, say B_1, B_2, are of degree 2 and correspond to $q_1(x)$ and $q_2(x)$, respectively. The rest, say T of degree 4, is the determinant of 4 by 4 matrix obtained from the four coefficients $A_0, \ldots, A_3 \in \mathbf{C}^2 \otimes \mathbf{C}^2$ of $A(x)$. There is one relation $T^2 = f_4(B_1, B_2)$. Hence $\mathrm{Proj}\, R^{SL(2)\times SL(2)}$ is a double cover of $\mathbf{P}^1$ as desired.

The moduli space $M_S(2, \mathcal{O}_S(1), 2)$ in Example 3 is constructed similarly. Let $\mathcal{A}lt_4$ be the affine space associated to the vector space $\bigoplus_{i=0}^5 (\bigwedge^2 \mathbf{C}^4)x_i$ and $\mathcal{A}lt_{4,2}$ the subscheme defined by the condition that $\sum_{i=0}^5 A_i x_i \in \mathcal{A}lt_4$ is of rank ≤ 2 everywhere on S_8. Then the invariant ring of the action of $SL(4)$ on $\mathcal{A}lt_{4,2}$ is generated by four elements B_1, B_2, B_3, T of degree 2, 2, 2, 6. There is one relation $T^2 = f_6(B_1, B_2, B_3)$ and $M_S(2, \mathcal{O}_S(1), 2)$, the projective spectrum $\mathrm{Proj}\, R^{SL(4)}$, is a double cover of $\mathbf{P}^2$ as described.

The moduli space of vector bundles on a surface was first constructed by Gieseker [6]. He took the Mumford's GIT quotient [19] of Grothendieck's Quot scheme [7] by PGL and used the Gieseker matrix to measure the stability of the

action. In the above construction, we take the quotient of $\mathcal{A}lt_{4,2}$, which is nothing but the affine variety of Gieseker matrices of suitable rank 2 vector bundles, by a general linear group $GL(4)$.

The Jacobian, or the Picard variety, of a curve is more fundamental. Weil [27] constructed $\mathrm{Pic}_g C$ as an algebraic variety using the symmetric product $\mathrm{Sym}^g C$ and showed its projectivity by Lefschetz' 3Θ theorem. Later Seshadri and Oda [24] constructed $\mathrm{Pic}_d C$ for arbitray d (over the same ground field as C) by also taking the GIT quotient of Quot schemes. The above constructions eliminate Quot schemes and the concept of linearization from those of Gieseker, Seshadri and Oda.

References

[1] K. Abe: A remark on the 2-dimensional moduli spaces of vector bundles on K3 surfaces, Math. Res. Letters, **7**(2000), 463–470.

[2] E. Arbarello, M. Cornalba, P. A. Griffiths, and J. Harris : *Geometry of algebraic curves, I*, Springer-Verlag, 1985.

[3] M. F. Atiyah : Vector bundles over an elliptic curve, Proc. London Math. Soc., **7**(1957), 414-452.

[4] T. Bridgeland: Equivalences of triangulated categories and Fourier-Mukai transformations, Bull. London Math. Soc., **31**(1999), 25-34.

[5] A. Căldăraru: Non-fine moduli spaces of sheaves on K3 surfaces, preprint.

[6] D. Gieseker: On the moduli of vector bundles on an algebraic surface, Ann. Math. **106**(1977), 45-60.

[7] A. Grothendieck : Techniques de construction et théorèm d'existence en géométrie algébrique, iV: Les schémas de Hilbert, Sem. Bourbaki, t. 13, 1960/61, n^o 221.

[8] V.A. Iskovskih : Fano 3-folds, II, Izv. Akad. Nauk SSSR,**42**(1978) : English translation, Math. USSR Izv. **12**(1978), 469–505.

[9] S. Mori and S. Mukai: The uniruledness of the moduli space of curves of genus 11, in 'Algebraic Geometry, Proceedings, Tokyo/Kyoto 1982', Series: Lecture Notes in Mathematics, vol. 1016, (M, Raynaud and T. Shioda eds.), Springer Verlag, 1983, 334–353.

[10] S. Mukai: Symplectic structure of the moduli space of sheaves on an abelian or $K3$ surface, Invent. Math., **77**(1984), 101–116.

[11] —: On the moduli space of bundles on $K3$ surfaces, I, in 'Vector Bundles on Algebraic Varieties ', Tata Institute of Fundamental Research, Bombay, 1987, 341–413.

[12] — : Curves, $K3$ surfaces and Fano manifolds of genus $\leq$ 10, in 'Algebraic Geometry and Commutative Algebra in honor of Masayoshi NAGATA', (H. Hijikata and H. Hironaka et al eds.), Kinokuniya, Tokyo, 1988, 367–377.

[13] — : Biregular classification of Fano threefolds and Fano manifolds of coindex 3, Proc. Nat. Acad. Sci., USA, **86** (1989), 3000–3002.

[14] — : New developments in the theory of Fano 3-folds: Vector bundle method and moduli problem, Sugaku, **47**(1995), 125–144.: English translation, Sugaku Expositions, to appear.

[15] —: Curves and K3 surfaces of genus eleven, in 'Moduli of Vector Bundles', Series: Pure and Applied Math., (Maruyama M. ed.), Mercel Dekker, New York, 1996, 189–197.
[16] — : Non-Abelian Brill-Noether theory and Fano 3-folds, Sugaku, **49**(1997), 1–24.: English translation, Sugaku Expositions, **14**(2001), 125–153.
[17] —: Duality of polarized K3 surfaces, in Proceedings of Euroconference on Algebraic Geometry, (K. Hulek and M. Reid ed.), Cambridge University Press, 1998, 107–122.
[18] —: *Moduli theory, I, II*, Iwanami Shoten, Tokyo, 1998, 2000: English translation, *An introduction to invariants and moduli*, to appear from Cambridge University Press.
[19] D. Mumford : *Geometric invariant theory*, Springer Verlag, 1965.
[20] P. E. Newstead : Stable bundles of rank 2 and odd degree over a curve of genus 2, Topology, **7**(1968), 205–215.
[21] V. V. Nikulin : On correspondences between surfaces of K3 type, Izv. Akad. Nauk SSSR Ser. Mat., **51** (1987), 402–411. English translation, Math. USSR Izv., **30**(1988), 375–383.
[22] D. Orlov : Equivalences of derived categories and K3 surfaces, J. Math. Sci. (New York), **84**(1997), 1361–1381.
[23] I. R. Shafarevich : Le théorème de Torelli pour les surfaces algébriques de type K3, Actes Congrès Intern. Math., Nice 1970, 413–417(1971).
[24] C.S. Seshadri and T. Oda : Compactifications of the generalized Jacobian Variety, Trans. Amer. Math. Soc., **253**(1979), 1–90.
[25] V.V. Shokurov : Smoothness of the general anticanonical divisor, Izv. Acad. Nauk SSSR **43**(1979), 430–441 : English translation, Math. USSR Izv. **14**(1980), 395–405.
[26] A. Tjurin : On intersection of quadrics, Russian Math. Survey **30**(1975), 51–105.
[27] A. Weil : *Variétés abéliennes et courbes algébriques*, Hermann, Paris, 1948.
[28] H. Weyl : *The classical groups*, Princeton Univ. Press, 1939.

ICM 2002 · Vol. II · 503–512

Three Questions in Gromov-Witten Theory

R. Pandharipande*

Abstract

Three conjectural directions in Gromov-Witten theory are discussed: Gorenstein properties, BPS states, and Virasoro constraints. Each points to basic structures in the subject which are not yet understood.

2000 Mathematics Subject Classification: 14N35, 14H10.
Keywords and Phrases: Gromov-Witten theory, Moduli of curves.

1. Introduction

Let X be a nonsingular projective variety over $\mathbb{C}$. Gromov-Witten theory concerns integration over $\overline{M}_{g,n}(X,\beta)$, the moduli space of stable maps from genus g, n-pointed curves to X representing the class $\beta \in H_2(X,\mathbb{Z})$. While substantial progress in the mathematical study of Gromov-Witten theory has been made in the past decade, several fundamental questions remain open. My goal here is to describe three conjectural directions:

(i) Gorenstein properties of tautological rings,
(ii) BPS states for threefolds,
(iii) Virasoro constraints.

Each points to basic structures in Gromov-Witten theory which are not yet understood. New ideas in the subject will be required for answers to these questions.

2. Gorenstein properties of tautological rings

The study of the structure of the entire Chow ring of the moduli space of pointed curves $\overline{M}_{g,n}$ appears quite difficult at present. As the principal motive is to understand cycle classes obtained from algebro-geometric constructions, we

*Department of Mathematics, Princeton University, Princeton, NJ 08544, USA. E-mail: rahulp@math.princeton.edu

may restrict inquiry to the system of *tautological rings*, $R^*(\overline{M}_{g,n})$. The tautological system is defined to be the set of smallest $\mathbb{Q}$-subalgebras of the Chow rings,

$$R^*(\overline{M}_{g,n}) \subset A^*(\overline{M}_{g,n}),$$

satisfying the following three properties:

(i) $R^*(\overline{M}_{g,n})$ contains the cotangent line classes $\psi_1, \ldots, \psi_n$ where

$$\psi_i = c_1(L_i),$$

the first Chern class of the ith cotangent line bundle.

(ii) The system is closed under push-forward via all maps forgetting markings:

$$\pi_* : R^*(\overline{M}_{g,n}) \to R^*(\overline{M}_{g,n-1}).$$

(iii) The system is closed under push-forward via all gluing maps:

$$\pi_* : R^*(\overline{M}_{g_1,n_1\cup\{*\}}) \otimes_{\mathbb{Q}} R^*(\overline{M}_{g_2,n_2\cup\{\bullet\}}) \to R^*(\overline{M}_{g_1+g_2,n_1+n_2}),$$

$$\pi_* : R^*(\overline{M}_{g_1,n_1\cup\{*,\bullet\}}) \to R^*(\overline{M}_{g_1+1,n_1}).$$

Natural algebraic constructions typically yield Chow classes lying in the tautological ring. See [7], [18] for further discussion.

Consider the following basic filtration of the moduli space of pointed curves:

$$\overline{M}_{g,n} \supset M^c_{g,n} \supset M^{rt}_{g,n} \supset C_{g,n}.$$

Here, $M^c_{g,n}$ denotes the moduli of pointed curves of compact type, $M^{rt}_{g,n}$ denotes the moduli of pointed curves with rational tails, and $C_{g,n}$ denotes the moduli of pointed curves with a fixed stabilized complex structure C_g. The choice of C_g will play a role below.

The tautological rings for the elements of the filtration are defined by the images of $R^*(\overline{M}_{g,n})$ in the associated quotient sequence:

$$R^*(\overline{M}_{g,n}) \to R^*(M^c_{g,n}) \to R^*(M^{rt}_{g,n}) \to R^*(C_{g,n}) \to 0. \tag{2.1}$$

Remarkably, the tautological rings of the strata are conjectured to resemble cohomology rings of compact manifolds.

A finite dimension graded algebra R is Gorenstein with socle in degree s if there exists an evaluation isomorphism,

$$\phi : R^s \xrightarrow{\sim} \mathbb{Q},$$

for which the bilinear pairings induced by the ring product,

$$R^r \times R^{s-r} \to R^s \xrightarrow{\phi} \mathbb{Q},$$

are nondegenerate. The cohomology rings of compact manifolds are Gorenstein algebras.

Conjecture 1. *The tautological rings of the filtration of* $\overline{M}_{g,n}$ *are finite dimensional Gorenstein algebras.*

The Gorenstein structure of $R^*(M_g)$ with socle in degree $g-2$ was discovered by Faber in his study of the Chow rings of M_g in low genus. The general conjecture is primarily motivated by Faber's original work and can be found in various stages in [5], [19], and [7].

The application of the conjecture to the stratum $C_{g,n}$ takes a special form due to the choice of the underlying curve C_g. The conjecture is stated for a nonsingular curve C_g defined over $\overline{\mathbb{Q}}$ or, alternatively, for the tautological ring in $H^*(C_{g,n}, \mathbb{Q})$. The tautological ring of $C_{g,n}$ in Chow is *not* Gorenstein for all C_g by recent results of Green and Griffiths.

Two main questions immediately arise if the tautological rings are Gorenstein algebras:

(i) Can the ring structure be described explicitly?
(ii) Are the tautological rings associated to embedded compact manifolds in the moduli space of pointed curves?

The tautological ring structures are implicitly determined by the conjectural Gorenstein property and the Virasoro constraints [10].

As the moduli space of curves may be viewed as a special case of the moduli space of maps, a development of these ideas may perhaps be pursued more fully in Gromov-Witten theory. It is possible to define a tautological ring for $\overline{M}_{g,n}(X,\beta)$ in the context of the virtual class by *assuming* the Gorenstein property, but no structure has been yet been conjectured. Again, the Virasoro constraints in principle determine the tautological rings.

3. BPS states for threefolds

Let X be a nonsingular projective variety over $\mathbb{C}$ of dimension 3. Let $\{\gamma_a\}_{a\in A}$ be a basis of $H^*(X,\mathbb{Z})$ modulo torsion. Let $\{\gamma_a\}_{a\in D_2}$ and $\{\gamma_a\}_{a\in D_{>2}}$ denote the classes of degree 2 and degree greater than 2 respectively. The Gromov-Witten invariants of X are defined by integration over the moduli space of stable maps (against the virtual fundamental class):

$$\langle \gamma_{a_1}, \ldots, \gamma_{a_n} \rangle_{g,\beta}^X = \int_{[\overline{M}_{g,n}(X,\beta)]^{vir}} \mathrm{ev}_1^*(\gamma_{a_1}) \ldots \mathrm{ev}_n^*(\gamma_{a_n}), \tag{3.1}$$

where ev_i is the ith evaluation map. As the moduli spaces are Deligne-Mumford stacks, the Gromov-Witten invariants take values in $\mathbb{Q}$.

Let $\{t_a\}$ be a set of variables corresponding to the classes $\{\gamma_a\}$. The Gromov-Witten potential $F^X(t,\lambda)$ of X may be written,

$$F^X = F^X_{\beta=0} + \tilde{F}^X, \tag{3.2}$$

as a sum of constant and nonconstant map contributions.

The constant map contribution $F^X_{\beta=0}$ may be further divided by genus:

$$F^X_{\beta=0} = F^0_{\beta=0} + F^1_{\beta=0} + \sum_{g\geq 2} F^g_{\beta=0}.$$

The genus 0 constant contribution records the classical intersection theory of X:

$$F^0_{\beta=0} = \lambda^{-2} \sum_{a_1,a_2,a_3\in A} \frac{t_{a_3}t_{a_2}t_{a_1}}{3!} \int_X \gamma_{a_1}\cup\gamma_{a_2}\cup\gamma_{a_3}.$$

The genus 1 constant contribution is obtained from a virtual class calculation:

$$F^1_{\beta=0} = \sum_{a\in D_2} t_a \langle\gamma_a\rangle^X_{g=1,\beta=0} = -\sum_{a\in D_2} \frac{t_a}{24}\int_X \gamma_a\cup c_2(X).$$

Similarly, the genus $g\geq 2$ contributions are

$$F^g_{\beta=0} = \langle 1\rangle^X_{g,\beta=0} = (-1)^g \frac{\lambda^{2g-2}}{2}\int_X (c_3(X) - c_1(X)\cup c_2(X))\cdot \int_{\overline{M}_g} \lambda^3_{g-1}.$$

The Hodge integrals which arise here have been computed in [6]:

$$\int_{\overline{M}_g} \lambda^3_{g-1} = \frac{|B_{2g}|}{2g}\frac{|B_{2g-2}|}{2g-2}\frac{1}{(2g-2)!},$$

where B_{2g} and B_{2g-2} are Bernoulli numbers. The constant map contributions to F^X are therefore completely understood.

The second term in (3.2) is the nonconstant map contribution:

$$\tilde{F}^X = \sum_{g\geq 0}\sum_{\beta\neq 0} F^g_\beta.$$

Since the virtual dimension of the moduli space $\overline{M}_g(X,\beta)$ is

$$\int_\beta c_1(X) + 3g - 3 + 3 - 3g = \int_\beta c_1(X),$$

the classes β satisfying $\int_\beta c_1(X) < 0$ do not contribute to the potential F^X. Therefore, $\tilde{F}^X$ may be divided into two sums:

$$\begin{aligned}\tilde{F}^X \quad = \quad & \sum_{g\geq 0} \sum_{\beta\neq 0,\ \int_\beta c_1(X)=0} F^g_\beta \\ & + \sum_{g\geq 0} \sum_{\beta\neq 0,\ \int_\beta c_1(X)>0} F^g_\beta.\end{aligned}$$

In case $\beta\neq 0$, we will write the series $F^g_\beta(t,\lambda)$ in the following form:

$$F^g_\beta = \lambda^{2g-2} q^\beta \sum_{n\geq 0}\frac{1}{n!}\sum_{a_1,\ldots,a_n\in D_{>2}} t_{a_n}\cdots t_{a_1}\langle\gamma_{a_1}\cdots\gamma_{a_n}\rangle^X_{g,\beta}.$$

The degree 2 variables $\{t_a\}_{a\in D_2}$ are formally suppressed in q via the divisor equation:

$$q^\beta = \prod_{a\in D_2} q_a^{\int_\beta \gamma_a}, \quad q_a = e^{t_a}.$$

Cohomology classes of degree 0 or 1 do not appear in nonvanishing Gromov-Witten invariants (3.1) for curve classes $\beta \neq 0$.

We will define new invariants $n_\beta^g(\gamma_{a_1},\ldots,\gamma_{a_n})$ for every genus g, curve class $\beta \neq 0$, and classes $\gamma_{a_1},\ldots,\gamma_{a_n}$. Our primary interest will be in the case where the following conditions hold:

(i) $\deg(\gamma_{a_i}) > 2$ for all i.
(ii) $n + \int_\beta c_1(X) = \sum_{i=1}^n \deg(\gamma_{a_i})$.

The invariants will be defined to satisfy the divisor equation (which allows for the extraction of degree 2 classes γ_a) and defined to vanish if degree 0 or 1 classes are inserted or if condition (ii) is violated. If $\int_\beta c_1(X) = 0$, then n_β^g is well-defined without cohomology insertions.

The new invariants $n_\beta^g(\gamma_{a_1},\ldots,\gamma_{a_n})$ are defined via Gromov-Witten theory by the following equation:

$$\begin{aligned}\tilde{F}^X \;=\; & \sum_{g\geq 0} \sum_{\beta\neq 0,\ \int_\beta c_1(X)=0} n_\beta^g\ \lambda^{2g-2} \sum_{d>0} \frac{1}{d}\left(\frac{\sin(d\lambda/2)}{\lambda/2}\right)^{2g-2} q^{d\beta} \\ & + \sum_{g\geq 0} \sum_{\beta\neq 0,\ \int_\beta c_1(X)>0} \sum_{n\geq 0} \frac{1}{n!} \sum_{a_1,\ldots,a_n\in D_{>2}} t_{a_n}\cdots t_{a_1} \\ & \qquad \cdot\, n_\beta^g(\gamma_{a_1},\cdots,\gamma_{a_n})\ \lambda^{2g-2}\left(\frac{\sin(\lambda/2)}{\lambda/2}\right)^{2g-2+\int_\beta c_1(X)} q^\beta.\end{aligned}$$

The above equation uniquely determines the invariants $n_\beta^g(\gamma_{a_1},\ldots,\gamma_{a_n})$.

Conjecture 2. *For all nonsingular projective threefolds X,*

(i) the invariants $n_\beta^g(\gamma_{a_1},\ldots,\gamma_{a_n})$ are integers,
(ii) for fixed β, the invariants $n_\beta^g(\gamma_{a_1},\ldots,\gamma_{a_n})$ vanish for all sufficiently large genera g.

If X is a Calabi-Yau threefold, the Gopakumar-Vafa conjecture is recovered [15], [16]. Here, the invariants n_β^g arise as BPS state counts in a study of Type IIA string theory on X via M-theory. The outcome is a physical deduction of the conjecture in the Calabi-Yau case.

Gopakumar and Vafa further propose a mathematical construction of the Calabi-Yau invariants n_β^g using moduli spaces of sheaves on X. The invariants n_β^g should arise as multiplicities of special representations of $\mathfrak{sl}_2$ in the cohomology of the moduli space of sheaves. The local Calabi-Yau threefold consisting of a curve C together with a rank 2 normal bundle N satisfying $c_1(N) = \omega_C$ should be the most basic case. Here the BPS states n_d^g should be found in the cohomology of an appropriate moduli space of rank d bundles on C. A mathematical development of

the proposed connection between integrals over the moduli of stable maps and the cohomology of the moduli of sheaves has not been completed. However, evidence for the program can be found both in local and global calculations in several cases [1], [20], [21].

The conjecture for arbitrary threefolds is motivated by the Calabi-Yau case together with the degeneracy calculations of [29]. Evidence can be found, for example, in the low genus enumerative geometry of $\mathbf{P}^3$ [9], [29]. If the conjecture is true, the invariants $n_\beta^g(\gamma_{a_1},\ldots,\gamma_{a_n})$ of $\mathbf{P}^3$ may be viewed as *defining* an integral enumerative geometry of space curves for all g and β. Classically the enumerative geometry of space curves does not admit a uniform description.

The conjecture does not determine the Gromov-Witten invariants of threefolds. A basic related question is to find some means to calculate higher genus invariants of Calabi-Yau threefolds. The basic test case is the quintic hypersurface in $\mathbf{P}^4$. There are several approaches to the genus 0 invariants of the quintic: Mirror symmetry, localization, degeneration, and Grothendieck-Riemann-Roch [2],[3], [8],[11],[23]. But, the higher genus invariants of the quintic are still beyond current string theoretic and geometric techniques. The best tool for the higher genus Calabi-Yau case, the holomorphic anomaly equation, is not well understood in mathematics. On the other hand, all the invariants of $\mathbf{P}^3$ may be in principle calculated by virtual localization [17].

4. Virasoro constraints

Let X be a nonsingular projective variety over $\mathbb{C}$ of dimension r. Let $\{\gamma_a\}$ be a basis of $H^*(X,\mathbb{C})$ homogeneous with respect to the Hodge decomposition,

$$\gamma_a \in H^{p_a,q_a}(X,\mathbb{C}).$$

The descendent Gromov-Witten invariants of X are:

$$\langle \tau_{k_1}(\gamma_{a_1})\ldots\tau_{k_n}(\gamma_{a_n})\rangle_{g,\beta}^X = \int_{[\overline{M}_{g,n}(X,\beta)]^{vir}} \psi_1^{k_1}\,\mathrm{ev}_1^*(\gamma_{a_1})\ldots\psi_n^{k_n}\,\mathrm{ev}_n^*(\gamma_{a_n}).$$

Let $\{t_k^a\}$ be a set of variables. Let $F^X(t,\lambda)$ be the generating function of the descendent invariants:

$$F^X = \sum_{g\geq 0}\lambda^{2g-2}\sum_{\beta\in H_2(X,\mathbb{Z})}q^\beta\sum_{n\geq 0}\frac{1}{n!}\sum_{\substack{a_1\ldots a_n\\ k_1\ldots k_n}} t_{k_n}^{a_n}\ldots t_{k_1}^{a_1}\langle\tau_{k_1}(\gamma_{a_1})\ldots\tau_{k_n}(\gamma_{a_n})\rangle_{g,\beta}^X.$$

The partition function Z^X is formed by exponentiating F^X:

$$Z^X = \exp(F^X). \tag{4.1}$$

We will now define formal differential operators $\{L_k\}_{k\geq -1}$ in the variables t_k^a satisfying the Virasoro bracket,

$$[L_k, L_\ell] = (k-\ell)L_{k+\ell}.$$

The definitions of the operators L_k will depend only upon the following three structures of $H^*(X,\mathbb{C})$:

(i) the intersection pairing $g_{ab} = \int_X \gamma_a \cup \gamma_b$,
(ii) the Hodge decomposition $\gamma_a \in H^{p_a,q_a}(X, \mathbb{C})$,
(iii) the action of the anticanonical class $c_1(X)$.

The formulas for the operators L_k are:

$$\begin{aligned} L_k = \quad & \sum_{m=0}^{\infty} \sum_{i=0}^{k+1} \Big([b_a+m]_i^k (C^i)_a^b \tilde{t}_m^a \partial_{b,m+k-i} \\ & \qquad + \frac{\hbar}{2}(-1)^{m+1} [b_a-m-1]_i^k (C^i)^{ab} \partial_{a,m}\partial_{b,k-m-i-1} \Big) \\ & + \frac{\lambda^{-2}}{2} (C^{k+1})_{ab} t_0^a t_0^b \\ & + \frac{\delta_{k0}}{48} \int_X \big((3-r)c_r(X) - 2c_1(X)c_{r-1}(X) \big), \end{aligned}$$

where the Einstein convention for summing over the repeated indices $a, b \in A$ is followed.

Several terms require definitions. For each class γ_a, a half integer b_a is obtained from the Hodge decomposition,

$$b_a = p_a + (1-r)/2.$$

The combinatorial factor $[x]_i^k$ is defined by:

$$[x]_i^k = e_{k+1-i}(x, x+1, \ldots, x+k),$$

where e_k is the kth elementary symmetric function. The matrix C_a^b is determined by the action of the anticanonical class,

$$C_a^b \gamma_b = c_1(X) \cup \gamma_a.$$

The indices of C are lowered and raised by the metric g_{ab} and its inverse g^{ab}. The terms $\tilde{t}_m^a$ and $\partial_{a,m}$ are defined by:

$$\begin{aligned} \tilde{t}_m^a &= t_m^a - \delta_{a0}\delta_{m1}, \\ \partial_{a,m} &= \partial/\partial t_m^a, \end{aligned}$$

where both are understood to vanish if $m < 0$.

Conjecture 3. *For all nonsingular projective varieties X, $L_k(Z^X) = 0$.*

The conjecture for varieties X with only (p,p) cohomology was made by Eguchi, Hori, and Xiong [4]. The full conjecture involves ideas of Katz. In case X is a point, the constraints specialize to the known Virasoro formulation of Witten's conjecture [22], [30] (see also [25]). After the point, the simplest varieties occur in two basic families: curves C_g of genus g and projective spaces $\mathbf{P}^n$ of dimension n. A proof of the Virasoro constraints for target curves C_g is presented in a sequence of papers [26], [27], [28]. Givental has recently proven the Virasoro

constraints for the projective spaces $\mathbf{P}^n$ [12], [13], [14]. The two families of varieties are quite different in flavor. Curves are of dimension 1, but have non-(p,p) cohomology, non-semisimple quantum cohomology, and do not always carry torus actions. Projective spaces cover all target dimensions, but have algebraic cohomology, semisimple quantum cohomology, and always carry torus actions.

The Virasoro constraints are especially appealing from the point of view of algebraic geometry as *all* nonsingular projective varieties are covered. While many aspects of Gromov-Witten theory may be more naturally pursued in the symplectic category, the Virasoro constraints appear to require more than a symplectic structure to define. For example, the bracket

$$[L_1, L_{-1}] = 2L_0,$$

depends upon formulas expressing the Chern numbers,

$$\int_X c_r(X), \quad \int_X c_1(X)c_{r-1}(X),$$

in terms of the Hodge numbers $h^{p,q}$ of X (see [24]]).

The Virasoro constraints may be a shadow of a deeper connection between the Gromov-Witten theory of algebraic varieties and integrable systems. In case the target is the point or the projective line, precise connections have been made to the KdV and Toda hierarchies respectively. The connections are proven by explicit formulas for the descendent invariants in terms of matrix integrals (for the point) and vacuum expectation in $\Lambda^{\frac{\infty}{2}}V$ (for the projective line) [22], [25], [27]. The extent of the relationship between Gromov-Witten theory and integrable systems is not known. In particular, an understanding of the surface case would be of great interest. Perhaps a link to integrable systems can be found in the circle of ideas involving Hilbert schemes of points, Heisenberg algebras, and Göttsche's conjectures concerning the enumerative geometry of linear series.

Finally, one might expect Virasoro constraints to hold in the context of Gromov-Witten theory relative to divisors in the target X. For the relative theory of 1-dimensional targets X, Virasoro constraints have been found and play a crucial role in the proof of the Virasoro constraints for the absolute theory of X [28].

References

[1] J. Bryan and R. Pandharipande, *BPS states of curves in Calabi-Yau 3-folds*, Geom. Topol. **5** (2001), 287–318.

[2] P. Candelas, X. de la Ossa, P. Green and L. Parkes, *A pair of Calabi-Yau manifolds as an exactly soluble superconformal field theory*, Nuclear Physics **B359** (1991), 21–74.

[3] T. Coates and A. Givental, *Quantum Riemann-Roch, Lefschetz, and Serre*, math.AG/0110142.

[4] T. Eguchi, K. Hori, and C.-S. Xiong, *Quantum cohomology and Virasoro algebra*, Phys. Lett. **B402** (1997), 71–80.

[5] C. Faber, *A conjectural description of the tautological ring of the moduli space of curves*, in Moduli of Curves and Abelian Varieties (The Dutch Intercity Seminar on Moduli), C. Faber and E. Looijenga, eds., 109–129, Aspects of Mathematics E33, Vieweg: Wiesbaden, 1999.
[6] C. Faber and R. Pandharipande, *Hodge integrals and Gromov-Witten theory*, Invent. Math. **139** (2000), 173–199.
[7] C. Faber and R. Pandharipande, *Logarithmic series and Hodge integrals in the tautological ring*, with an appendix by Don Zagier, Michigan Math. J. **48** (2000), 215–252.
[8] A. Gathmann, *Relative Gromov-Witten invariants and the mirror formula*, math.AG/0202002.
[9] E. Getzler, *Intersection theory on $\overline{M}_{1,4}$ and elliptic Gromov-Witten invariants*, J. Amer. Math. Soc. **10** (1997), 973–998.
[10] E. Getzler and R. Pandharipande, *Virasoro constraints and the Chern classes of the Hodge bundle*, Nucl. Phys. **B530** (1998), 701–714.
[11] A. Givental, *Equivariant Gromov-Witten invariants*, Int. Math. Res. Notices **13** (1996), 613–663.
[12] A. Givental, *Semisimple Frobenius structures at higher genus*, math.AG/0008067.
[13] A. Givental, *Gromov-Witten invariants and quantization of quadratic hamiltonians*, math.AG/0108100.
[14] A. Givental, in preparation.
[15] R. Gopakumar and C. Vafa, *M-theory and topological strings I*, hep-th/9809187.
[16] R. Gopakumar and C. Vafa, *M-theory and topological strings II*, hep-th/9812127.
[17] T. Graber and R. Pandharipande, *Localization of virtual classes*, Invent. Math. **135** (1999), 487–518.
[18] T. Graber and R. Pandharipande, *A non-tautological algebraic class on $\overline{M}_{2,22}$*, math.AG/0104057.
[19] R. Hain and E. Looijenga, *Mapping class groups and moduli spaces of curves*, in Proceedings of Symposia in Pure Mathematics: Algebraic Geometry Santa Cruz 1995, J. Kollár, R. Lazarsfeld, D. Morrison, eds., Volume 62, Part 2, 97–142.
[20] S. Hosono, M.-H. Saito, and A. Takahashi, *Holomorphic anomaly equation and BPS state counting of rational elliptic surface*, Adv. Theor. Math. Phys. **1** (1999), 177–208.
[21] S. Katz, A. Klemm, and C. Vafa, *M-theory, topological strings, and spinning black holes,* hep-th/9910181
[22] M. Kontsevich, *Intersection theory on the moduli space of curves and the matrix Airy function*, Comm. Math. Phys. **147** (1992), 1–23.
[23] B. Lian, K. Liu, and S.-T. Yau, *Mirror principle I*, Asian J. Math. **1** (1997), 729–763.
[24] A. Libgober and J. Wood, *Uniqueness of the complex structure on Kähler manifolds of certain homology types*, J. Diff. Geom. **32** (1990), 139–154.
[25] A. Okounkov and R. Pandharipande, *Gromov-Witten theory, Hurwitz numbers,*

and matrix models, I, math.AG/0101147 .
[26] A. Okounkov and R. Pandharipande, *Gromov-Witten theory, Hurwitz theory, and completed cycles*, math.AG/0204305.
[27] A. Okounkov and R. Pandharipande, *The equivariant Gromov-Witten theory of* $\mathbf{P}^1$, in preparation.
[28] A. Okounkov and R. Pandharipande, *The Gromov-Witten theory of target curves*, in preparation.
[29] R. Pandharipande, *Hodge integrals and degenerate contributions*, Comm. Math. Phys. **208** (1999), 489–506.
[30] E. Witten, *Two dimensional gravity and intersection theory on moduli space*, Surveys in Diff. Geom. **1** (1991), 243–310.

ICM 2002 · Vol. II · 513–524

Update on 3-folds

Miles Reid*

Abstract

The familiar division of compact Riemann surfaces into 3 cases

$$g = 0, \quad g = 1 \quad \text{and} \quad g \geq 2$$

corresponds to the well known trichotomy of spherical, Euclidean and hyperbolic non-Euclidean plane geometry. *Classification* aims to treat all projective algebraic varieties in terms of this trichotomy; the model is Castelnuovo and Enriques' treatment of surfaces around 1900 (reworked by Kodaira in the 1960s). The canonical class of a variety may not have a definite sign, so we usually have to beat it up before the trichotomy applies, by a minimal model program (MMP) using contractions, flips and fibre space decompositions. The classification of 3-folds was achieved by Mori and others during the 1980s.

New results over the last 5 years have added many layers of subtlety to higher dimensional classification. The study of 3-folds also yields a rich crop of applications in several different branches of algebra, geometry and theoretical physics. My lecture surveys some of these topics.

2000 Mathematics Subject Classification: 14E30, 14J30, 14J32, 14J35, 14J45, 14J81.
Keywords and Phrases: Mori theory, Minimal model program, Classification of varieties, Fano 3-folds, Birational geometry.

1. Popular introduction: the great trichotomy

A trichotomy is a logical division into three cases, where we expect to win something in each case. The cases here are similar to the "much too small, just right, much too big" of Goldilocks and the Three Bears, or the geometric division of conic sections into ellipse, parabola and hyperbola due to Appollonius of Perga (200 BC), or the cosmological question of whether the universe contracts again into a big crunch, tends to an asymptotic state or continues expanding exponentially.

*Mathematics Institute, University of Warwick, Coventry CV4 7AL, England, UK. E-mail: miles@maths.warwick.ac.uk

1.1. Euclidean and non-Euclidean geometry

Euclid's famous parallel postulate (c. 300 BC) states that

> if a line falls on two lines, with interior angles on one side adding to $< 180°$, the two lines, if extended indefinitely, meet on the side on which the angles add to $< 180°$.

We are in plane geometry, assumed homogeneous so that any construction involving lines, distances, angles, triangles and so on can be carried out at any point and in any orientation with the same effect. In this context the great trichotomy is the observation, probably due originally to Omar Khayyam (11th c.), Nasir al-Din al-Tusi (13th c.) and Gerolamo Saccheri (1733), that two other cases besides Euclid's are logically coherent (see Figure 1). In spherical geometry, the two lines meet

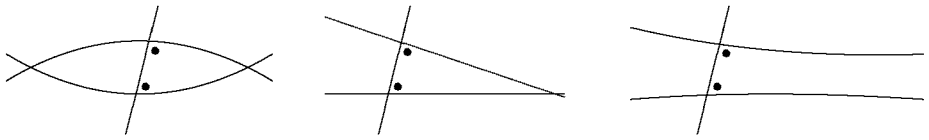


Figure 1: The parallel postulate

on *both* sides whatever the angles, whereas in hyperbolic non-Euclidean geometry, the two lines may diverge even though the angle sum is $< 180°$. Whether lines eventually meet is a long-range question, but it reflects the local *curvature* of the geometry.

1.2. Gauss and Riemann on differential geometry

A local surface S in 3-space is *positively curved* if all its sections bend in the same direction like the top of a sphere (see Figure 2). S is *flat* (or developable) if

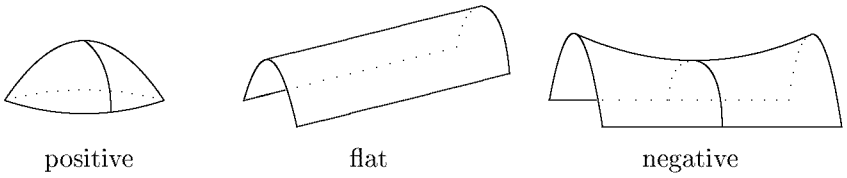


Figure 2: Local curvature

it is straight in one direction like a cylinder, and *negatively curved* if its sections bend in opposite directions like a saddle or Pringle's chip. Gauss in his *Theorema Egregium* (1828) and Riemann in his Habilitationsschrift (1854) found that curvature is intrinsic to the local distance geometry of S, independent of how S sits in 3-space: living on a sphere S of radius R, we can measure the perimeter of a disc of radius r, which is $2\pi(\sin \frac{r}{R})R$, always less than the Euclidean value $2\pi r$. If we lived

in the hyperbolic plane, the perimeter of a disc of radius r would be $2\pi(\sinh \frac{r}{R})R$, bigger than the Euclidean value, and growing exponentially with r.

Riemann in particular generalised Gauss' ideas on surfaces to a space given locally by an n-tuple $(x_1, \dots, x_n)$ of real parameters (a "many-fold extended quantity" or *manifold*), with *distance* arising from a local arc length $\mathrm{d}s$ given by a quadratic form $\mathrm{d}s^2 = \sum g_{ij}\mathrm{d}x_i\mathrm{d}x_j$. The curvature is then a function of the second derivatives of the metric function g_{ij}. Riemann's differential geometry works with manifolds that are not homogeneous, e.g., having positive, zero, or negative curvature at different points. It was a key ingredient in Einstein's general relativity (1915), which treats gravitation as curvature of space-time.

1.3. Riemann surfaces

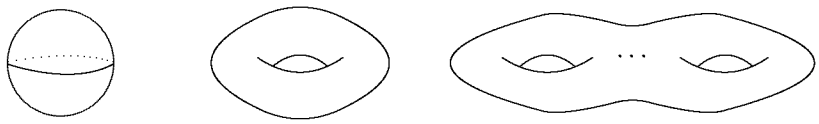
The story moves on from real manifolds (e.g., surfaces depending on 2 real variables) to Riemann surfaces, parametrised instead by a single complex variable. The point here is Cauchy's discovery (c. 1815) that differentiable functions of a complex variable are better behaved than real functions, and much more amenable to algebraic treatment. Riemann discovered that a compact Riemann surface C has an embedding $C \hookrightarrow \mathbb{P}^N_{\mathbb{C}}$ into complex projective space whose image is defined by a set of homogeneous polynomial equations.

A projective algebraic curve $C \subset \mathbb{P}^N_{\mathbb{C}}$ is *nonsingular* if at every point $P \in C$ we can choose $N-1$ local equations $f_1, \dots, f_{N-1}$ so that the Jacobian matrix $\frac{\partial f_i}{\partial x_j}$ has maximal rank $N-1$. It follows from the implicit function theorem that one of the linear coordinates $z = z_1$ of $\mathbb{P}^N$ can be chosen as a local analytic coordinate on C. In other words, a compact Riemann surface is analytically isomorphic to a nonsingular complex projective curve.

1.4. The genus of an algebraic curve

The *canonical class* $K_C = \Omega^1_C = T^*_C$ of a curve C is the holomorphic line bundle of 1-forms on C; it has transition functions on $U \cap U'$ the Jacobian of the coordinate change $\frac{\partial z'}{\partial z}$, where z, z' are local analytic coordinates on U, U'. If z is a rational function on C that is an analytic coordinate on an open set $U \subset C$ then a 1-form on U is $f(z)\mathrm{d}z$ with f a regular function on U. That is, $\Omega^1_C = \mathcal{O} \cdot \mathrm{d}z$, or $\mathrm{d}z$ is a basis of Ω^1_C on U.

The genus $g(C)$ can be defined in several ways: topologically, a compact Riemann surface is a sphere with g handles (see Figure 3). It has Euler number



$g = 0$, sphere $\qquad$ $g = 1$, torus $\qquad$ $g \geq 2$, general type

Figure 3: The genus of a Riemann surface

$e(C) = 2 - 2g$, which equals $\deg T_C$. The most useful formula for our purpose is

$\deg K_C = 2g-2$. We see that

$$K_C < 0 \iff g = 0, \qquad K_C = 0 \iff g = 1, \qquad K_C > 0 \iff g \geq 2.$$

This trichotomy is basic for the study of a curve C from every point of view, including topology, differential geometry, complex function theory, moduli, all the way through to algebraic geometry and Diophantine number theory. To relate this briefly to curvature as discussed in Section 1.2, for an arbitrary Riemannian metric, the *average* value of curvature over C equals $-\deg K_C$ by the Gauss–Bonnet theorem; moreover, by the Riemann mapping theorem, there exists a metric on C in the conformal class of the complex structure with constant positive, zero or negative curvature in the three cases.

2. Classification of 3-folds

The great trichotomy also drives classification in higher dimensions. The meaning of "higher dimensions" is time-dependent: dim 2 was worked out around 1900 by Castelnuovo and Enriques, dim 3 during the 1980s by Mori and others, and dim 4 is just taking off with Shokurov's current work. I concentrate on dim 3, where these issues first arose systematically.

2.1. Preliminaries: the canonical class K_X

An n-dimensional projective variety X can be embedded $X \hookrightarrow \mathbb{P}^N_{\mathbb{C}}$, and is given there by homogeneous polynomial equations; nonsingular means that at every point $P \in X$, we can choose $N-n$ of the defining equations so that the Jacobian matrix $\frac{\partial f_i}{\partial x_j}$ has rank $N-n$, with n linear coordinates of $\mathbb{P}^N$ providing local analytic coordinates on X.

The canonical class of X is $K_X = \Omega^n_X = \bigwedge^n \Omega^1_X$. It has many interpretations: it is the line bundle obtained as the top exterior power of the holomorphic cotangent bundle; it has transition functions on $U \cap U'$ the Jacobian determinant $\det \left| \frac{\partial x'_i}{\partial x_j} \right|$, where $\underline{x}, \underline{x}'$ are systems of local analytic coordinates on open sets $U, U' \subset X$; its sections are holomorphic n-forms; at a nonsingular point $P \in X$, its sections are generated by the holomorphic volume form $\mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_n$, so that $\Omega^n_X = \mathcal{O}_X \cdot \mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_n$.

For MMP to work in dim ≥ 3, we are eventually forced to allow certain mild singularities. The theory in dim 3 is now standard and not very hard (see [YPG] and compare the foreword to [CR]). We always insist that the first Chern class of K_X restricted to the nonsingular locus $X^0 \subset X$ comes from an element of $H^2(X, \mathbb{Q})$, that I continue to denote by K_X. This ensures that the pullback f^*K_X by a morphism $f\colon Y \to X$ is defined, together with the intersection number $K_X C$ with every curve $C \subset X$ (obtained by evaluating $K_X \in H^2$ against the class $[C] \in H_2(X, \mathbb{Q})$). Note that $-K_X C$ is an integral or *average* value of Ricci curvature (a 2-form) calculated over a 2-cycle $[C]$ corresponding to a holomorphic curve; we have taken several steps back from varieties of constant curvature suggested by the colloquial pictures in Section 1.

In many contexts, the canonical class of a variety is closely related to the *discrepancy divisor.* If $f\colon Y \to X$ is a birational morphism, its discrepancy Δ_f is defined by $K_Y = f^*K_X + \Delta_f$; if X and Y are nonsingular this is the divisor of zeros $\Delta_f = \operatorname{div}\left(\det\left|\frac{\partial x_i}{\partial y_j}\right|\right)$ of the Jacobian determinant of f, or its appropriate generalisation if X and Y are singular. Since the components of Δ_f are exceptional, it follows that if $\Delta_f > 0$, then there exists a component E of Δ_f such that $K_Y C < 0$ for almost every curve $C \subset E$. It is known that every section $s \in H^0(Y, nK_Y)$ vanishes along Δ_f for every $n \geq 0$. A morphism f is *crepant* if $\Delta_f = 0$; then $K_Y = f^*K_X$, so that K_Y is numerically zero relative to f.

2.2. The trichotomy: $K_X < 0$, $K_X = 0$ or $K_X > 0$?

The naive section heading is misleading: K_X may have "different sign" at different points of X and in different directions. The aim is not to apply the trichotomy to X itself, but to modify it first to a variety X' by a MMP. We need to be more precise; we say that K_X is *nef* or numerically nonnegative if $K_X C \geq 0$ for every $C \subset X$ (nef is an acronym for *numerically eventually free* – we hope that $|nK_X|$ is a free linear system for some $n > 0$). As we saw at the end of Section 2.1, a discrepancy divisor $\Delta_f > 0$ for a birational morphism $f\colon Y \to X$ is a local obstruction to the nefdom of K_Y. Mori theory (or the MMP) is concerned with the case that K_X is not nef.

2.3. Results of MMP for 3-folds

The *Mori category* consists of (quasi-)projective n-folds X with $\mathbb{Q}$-factorial terminal singularities; see [YPG] for details. For X in the Mori category, an *elementary contraction* is a morphism $\varphi\colon X \to X_1$ such that

(i) X_1 is a normal variety and φ has connected fibres.
(ii) All curves $C \subset X$ contracted by φ have classes in a single ray in $H_2(X, \mathbb{R})$, and $K_X C < 0$. This implies that $-K_X$ is relatively ample.

An elementary contraction $X \to S$ with $\dim S < \dim X$ is a *Mori fibre space* (Mfs). The case to bear in mind is when $S = \text{pt.}$; then (ii) implies that $-K_X$ is ample and $\rho(X) = \operatorname{rank}\operatorname{Pic} X = 1$, that is, X is a *Fano 3-fold* with $\rho = 1$. If $\dim S = \dim X - 1$ then $X \to S$ is a *conic bundle.*

Theorem 1 (see for example [KM])] *An elementary contraction exists if and only if K_X is not nef. For any 3-fold X in the Mori category there is a chain of birational transformations*

$$X \dashrightarrow X_1 \dashrightarrow \cdots \dashrightarrow X_n = X'$$

where (1) each step $X_i \dashrightarrow X_{i+1}$ is an elementary divisorial contraction or flip of the Mori category, and (2) the final object X' either *has $K_{X'}$ nef,* or *has a Mfs structure $X' \to S$.*

Each birational step $X_i \dashrightarrow X_{i+1}$ removes a subvariety of X on which K_X is negative. A *divisorial contraction* contracts an irreducible surface in X to a curve

or a point. A *flip* is a surgery operation that cuts out a finite number of curves in X_i on which K is negative, replacing them with curves on which K is positive. At the end of the MMP comes the dichotomy: either $K_{X'}$ is nef, or $-K_{X'}$ is ample on a global structure of X'.

The main theorem on varieties with K_X nef is the existence of an Iitaka–Kodaira fibration $X \to Y$, with fibres the curves $C \subset X$ with $K_X C = 0$. This gives a natural case division according to $\dim Y$. The extreme cases are Calabi-Yau varieties (CY), where $K_X = 0$, and varieties of general type, where $X \to Y$ is birational to a canonical model Y having canonical singularities and ample K_Y.

This takes my story up to around 1990; for more details, see Kollár and Mori [KM] or Matsuki [M].

3. Lots of recent progress

3.1. Extension of MMP to dimension 4

Already from the mid 1980s, it was understood that the MMP could in large parts be stated in all dimensions as a string of conjectures (or the log MMP, where we proceed in like manner, but directed by a log canonical class $K_X + D$). The difficult parts in $\dim \geq 3$ are the existence of flips (or log flips), and the termination of a chain of flips. Recent work of Shokurov [Sh] has established the existence of log flips in dim 4; the key idea is the reduction to prelimiting flips, already prominent in Shokurov's earlier work (see [FA], Chapter 18).

3.2. Rationally connected varieties

A variety X is *rational* if it is birationally equivalent to $\mathbb{P}^n$. That is, there are dense Zariski open sets $X_0 \subset X$ and $U \subset \mathbb{P}^n$, and an isomorphism $X_0 \cong U$ such that both φ and φ^{-1} are given by rational maps. In other words, X has a one-to-one parametrisation by rational functions. By analogy with curves and surfaces, one might hope that rational varieties have nice characterisations, and that rationality behaves well under passing to images or under deformation. Unfortunately, in $\dim \geq 3$, our experience is that this is not the case, and we are obliged to give up on the question of rationality.[1]

However, it turns out that the notion of *rationally connected* variety developed independently by Campana and by Kollár, Miyaoka and Mori is a good substitute. X is rationally connected if there is a rational curve through any two points $P, Q \in X$. See [Ca], [KMM], [Ko] and [GHS] for developments of this notion.

[1] This is of course exaggerated. Rationality itself remains the major issue in many contexts, in particular the rationality of GIT quotients. Iskovskikh's conjectured rationality criterion for conic bundles remains one of the driving forces of 3-fold birational geometry. Thanks to Slava Shokurov for reminding me of this important point.

3.3. Explicit classification results for 3-folds

Section 2.3 discussed the *Mori category* and its elementary contractions. The *explicit classification* manifesto of the foreword of [CPR] calls for the abstract definitions and existence results to be translated into practical lists of normal forms. The ideal result here is Mori's theorem [YPG], Theorem 6.1, that classifies 3-fold terminal singularities into a number of families; these relate closely to cyclic covers between Du Val singularities, and deform to varieties having only the terminal cyclic quotient singularities $\frac{1}{r}(1, a, -a)$.

To complete our grasp of Mori theory, we hope for explicit classification results in this style for divisorial contractions, flips and Mfs. The last few years have seen remarkable progress by Kawakita [Ka1], [Ka2] on divisorial contractions to points. A guiding problem in this area was Corti's 1994 conjecture ([Co2], p. 283) that every Mori divisorial contraction $\varphi\colon X \to Y$ to a nonsingular point $P \in Y$ is a $(1, a, b)$ weighted blowup. Kawakita proved this, and went on to classify explicitly the divisorial contractions to compound Du Val singularities of type A. There are also results of Tziolas on contractions of surfaces to curves. For progress on Mfs see Section 4.3.

3.4. Calabi-Yau 3-folds and mirror symmetry

A CY manifold X is a Kähler manifold with $K_X = 0$, usually assumed simply connected, or at least having $H^1(\mathcal{O}_X) = 0$. A popular recipe for constructing CY 3-folds is due to Batyrev, based on resolving the singularities of toric complete intersections. This gives some 500,000,000 families of CY 3-folds, so much more impressive than a mere infinity (see the website [KS]). There are certainly many more; I believe there are infinitely many families, but the contrary opinion is widespread, particularly among those with little experience of constructing surfaces of general type.

Calabi-Yau 3-folds are the scene of exciting developments related to the Strominger-Yau-Zaslow special Lagrangian approach to mirror symmetry. For lack of space, I refer to Gross [Gr] for a recent discussion.

3.5. Resolution of orbifolds and McKay correspondence

Klein around 1870 and Du Val in the 1930s studied quotient singularities $\mathbb{C}^2/G$ for finite groups $G \subset \mathrm{SL}(2, \mathbb{C})$. Du Val characterised them as singularities that "do not affect the condition of adjunction", that is, as surface canonical singularities. Quotient singularities $\mathbb{C}^3/G$ by finite subgroups $G \subset \mathrm{SL}(3, \mathbb{C})$ were studied by many authors around 1990; they proved case-by-case that a crepant resolution exists, and that its Euler number is equal to the number of conjugacy classes of G, as predicted by string theorists. The McKay correspondence says that the geometry of the crepant resolution of $\mathbb{C}^3/G$ can be described in terms of the representation theory of G. This has now been worked out in a number of contexts; see my Bourbaki talk [Bou].

3.6. The derived category as an invariant of varieties

The derived category $D(\mathcal{A})$ of an Abelian category $\mathcal{A}$ was introduced by Grothendieck and Verdier in the 1960s as a technical tool for homological algebra. A new point of view emerged around 1990 inspired by results of Beilinson and Mukai: for a projective nonsingular variety X over $\mathbb{C}$, write $D(X)$ for the bounded derived category of coherent sheaves on X; following Bondal and Orlov, we consider $D(X)$ *up to equivalence of* $\mathbb{C}$*-linear triangulated category* as an invariant of X, somewhat like a homology theory; the Grothendieck group $K_0(X)$ is a natural quotient of $D(X)$.

The derived category $D(X)$ is an enormously complicated and subtle object (already for $\mathbb{P}^2$); in this respect it is like the Chow groups, that are usually infinite dimensional, and contain much more information than anyone could ever use. Despite this, there are contexts, usually involving moduli constructions, in which "tautological" methods give equivalences of derived categories between $D(X)$ and $D(Y)$. An example is the method of [BKR] that establishes the McKay correspondence on the level of derived categories by Fourier–Mukai transform. There is no such natural treatment for the McKay correspondence in ordinary (co-)homology (see [Cr]).

The following conjectural discussion is based on ideas of Bondal, Orlov and others, as explained by Bridgeland (and possibly only half-understood by me). As I said, classification divides up all varieties into $K > 0$, $K = 0$, $K < 0$ and constructions made from them. Current work with $D(X)$ assumes that X is nonsingular, but I ignore this technical point. There must be some sense in which the derived category of a variety with $K < 0$ is "small" or "discrete"; for example, a semi-orthogonal sum of discrete pieces arising from smaller dimension. A contraction of the MMP should break off a little $K < 0$ semi-orthogonal summand; for nonsingular blowups, this is known [O], and also for certain flips [K]. For a variety X with $K = 0$, we expect $D(X)$ to have enormous symmetry, like a K3 or CY 3-fold; and for a variety with $K > 0$, $D(X)$ should be very infinite but rigid and indecomposable. Bondal and Orlov [BO] have proved that $D(X)$ determines X uniquely if $\pm K_X$ is ample, but as far as I know, they have not established a qualitative difference between the two cases.

Right up to Kodaira's work on surfaces in the 1960s, minimal models were seen in terms of tidying away -1-curves to make a really neat choice of model in a birational class, that eventually turns out to be unique. In contrast, starting from around 1980, the MMP in Mori theory sets itself the direct aim of making K nef if possible. Derived categories give us a revolutionary new aim: each step of the MMP chops off a little semi-orthogonal summand of $D(X)$.

4. Fano 3-folds: biregular and birational geometry

4.1. The Sarkisov program

The modern view of MMP and classification of varieties is as a *biregular* theory: although we classify varieties up to birational equivalence, the aims and the methods are stated in biregular terms. Beyond the MMP, the main birational problems are the following:

(1) If X is birational to a Mfs as in Theorem 1, then *in how many different ways* is it birational to a Mfs?
(2) Can we decide when two Mfs are birationally equivalent?
(3) Can we determine the group of birational selfmaps of a Mfs?

The Sarkisov program gives general answers to these questions, at least in principle. It untwists any birational map $\varphi\colon X \dashrightarrow Y$ between the total spaces of two Mfs X/S and Y/T as a chain of links, generalising Castelnuovo's famous treatment of birational maps of $\mathbb{P}^2$. A Sarkisov link of Type II consists of a Mori divisorial extraction, followed by a number of antiflips, flops and flips (in that order), then a Mori divisorial contraction.

More generally, the key idea is always to reduce to a 2-*ray game* in the Mori category (see [Co2], 269–272). By definition of Mfs, we have $\rho(X/S) = 1$, but for a 2-ray game we need a contraction $X' \to S'$ with $\rho(X'/S') = 2$. A Sarkisov link starts in one of two ways (depending on the nature of the map φ we are trying to untwist): either blow X up by a Mori extremal extraction $X' \to X$ and leave $S' = S$; or find a contraction $S \to S'$ of S so that $\rho(X/S') = 2$ and leave $X = X'$. In either case, the Mori cone of the new X'/S' is a wedge in $\mathbb{R}^2$ with a marked Mori extremal ray, and we can play a 2-ray game that contracts the other ray, flipping it whenever it defines a small contraction. It is proved that, given $\varphi\colon X \to Y$, one or other of these games can be played, and the link ends as it began in a Mori divisorial contraction or a change of Mfs structure, making four types of links. Each link decreases a (rather complicated) invariant of φ, and it is proved that a chain of links terminates. See [Co] and Matsuki [M] for details.

4.2. Birational rigidity

While the Sarkisov program factors birational maps as a chain of links that are elementary in some categorical sense, an explicit description of general links is still a long way off. To obtain generators of the Cremona group of $\mathbb{P}^3$ would involve classifying every Mfs X/S that is rational, and every Sarkisov link between these; for the time being, this is an impossibly large problem. There is, however, a large and interesting class of Mfs for which there are rather few Sarkisov links.

A Mori fibre space $X \to S$ is *birationally rigid* if for any other Mfs $Y \to T$, a birational map $\varphi\colon X \dashrightarrow Y$ can only exist if it lies over a birational map $S \dashrightarrow T$ such that X/S and Y/T have isomorphic general fibres (but φ need not induce an isomorphism of the general fibres – this is a tricksy definition). If $S = \text{pt.}$, so that X is a Fano variety with $\rho(X) = 1$, the condition means that the only Mfs Y/T birational to X is $Y \cong X$ itself. For example, $\mathbb{P}^2$ is not rigid, since it is birational to all the scrolls $\mathbb{F}_n$. Following imaginative but largely non-rigorous work of Fano in the 1930s, Iskovskikh and Manin proved in 1971 that a nonsingular quartic 3-fold $X_4 \subset \mathbb{P}^4$ is birationally rigid. This proof has since been simplified and reworked

by many authors. The main result of [CPR] is that a general element X of any of the *famous 95* families of Fano hypersurface $X_d \subset \mathbb{P}(1, a_1, \ldots, a_4)$ is likewise birationally rigid.

It is interesting to take a result of Corti and Mella [CM] as an example going beyond the framework of [CPR]. The codim 2 complete intersection $X_{3,4} \subset \mathbb{P}^5(1,1,1,1,2,2)$ is a Fano 3-fold; write $x_1, \ldots, x_4, y_1, y_2$ for homogeneous coordinates and $f_3 = g_4 = 0$ for the equations of $X_{3,4}$. By a minor change of coordinates, I can assume that $g_4 = y_1 y_2 + g'(x_1, \ldots, x_4)$. Then $X_{3,4}$ has $2 \times \frac{1}{2}(1,1,1)$ quotient singularities at the y_1, y_2 coordinate points. [CM] shows that blowing up either of these point leads to a Sarkisov link

$$\begin{array}{ccccc} X_{3,4} & \dashrightarrow & Y_5 & \dashleftarrow & Z_4 \\ \cap & & \cap & & \cap \\ \mathbb{P}^5(1^4, 2^2) & & \mathbb{P}^4(1^4, 2) & & \mathbb{P}^4 \end{array} \tag{4.1}$$

Here the midpoint Y_5 of the link is a general quintic containing the plane $\Pi = \mathbb{P}^2$, say given by $\Pi : (x_4 = y_1 = 0)$. Thus $Y_5 : (A_4 x_4 - B_3 y_1 = 0)$, where A_4, B_3 are quartic and cubic; note that Y_5 itself is not in the Mori category, because it is not factorial. We obtain $X_{3,4}$ by adding $y_2 = \frac{A_4}{y_1} = \frac{B_3}{x_4}$ to its homogeneous coordinate ring, and Z_4 by adding $x_0 = \frac{y_1}{x_4} = \frac{A_4}{B_3}$.

This example makes several points: $X_{3,4}$ and Z_4 are both Mori Fano 3-folds with $\rho = 1$. They are not birationally rigid, since they are birational to one another. [CM] proves that they are not birational to any Mfs other than $X_{3,4}$ and Z_4, so they form a *bi-rigid pair*. $X_{3,4}$ is general in its family, whereas Z_4 has in general a double point locally isomorphic to $x^2 + y^2 + z^3 + t^3$. This is a new kind of phenomenon that arises many times as soon as we go beyond the Fano hypersurfaces.

4.3. Explicit classification of Fano 3-folds

The anticanonical ring $R(X, -K_X) = \bigoplus H^0(-nK_X)$ of a Fano 3-fold X is a Gorenstein ring. Choosing a minimal set of homogeneous generators $x_0, \ldots, x_N$ of R with $\operatorname{wt} x_i = a_i$ defines an embedding $X \hookrightarrow \mathbb{P}(a_0, \ldots, a_N)$ as a projectively normal variety. The *codimension* of X is its codimension $N - 3$ in this embedding. If $N \leq 3$ the equations defining X are well understood, and we can describe X explicitly. For example, Altınok [Al] gives 69 families of Fano 3-folds whose general element has anticanonical ring of codim 3, given by the 4×4 Pfaffians of a 5×5 matrix, that is, a section of a weighted Grassmannian $\mathrm{wGr}(2,5)$ in the sense of [CR2].

The paper [ABR] explains how to use the formulas of [YPG] and the ideas of [Al] to make a computer database that includes all possible Hilbert series for $R(X, -K_X)$. In most cases the rings themselves can be studied by projection methods, as described in [Ki], in fact usually by projections of the simplest type. In other words, as in (4.1), we can make a weighted blowup $Y \to X$ of a terminal quotient singularity of X of type $\frac{1}{r}(1, a, -a)$. If we know $R(Y, -K_Y)$ and the ideal of the blown up $\mathbb{P}(1, a, -a)$ in it, we can reconstruct X by Kustin–Miller unprojection [PR]. Takagi's examples in [Ki], 6.4 and 6.8 is a warning that this process

is entertaining and nontrivial: there are *two different* families of Fano 3-folds in codim 4 with the same Hilbert series, obtained by unprojections that are numerically identical, and that differ only in the way that their unprojection planes embed $\Pi = \mathbb{P}^2 \hookrightarrow \mathrm{wGr}(2,5)$ in the weighted Grassmannian. These are the *Tom and Jerry unprojections* of [Ki], Section 8. The K3 surface sections of the two families form a single unobstructed family, but their extension to Fano 3-folds break up into two families; this is reminiscent of the extension-deformation theory of the del Pezzo surface of degree S_6, which has both $\mathbb{P}^2 \times \mathbb{P}^2$ and $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$ as extensions.

References

[Al] S. Altınok, Graded rings corresponding to polarised K3 surfaces and $\mathbb{Q}$-Fano 3-folds, Univ. of Warwick Ph.D. thesis, Sep. 1998, vii+93, get from www.maths.warwick.ac.uk/∼miles/doctors/Selma.

[ABR] S. Altınok, G. Brown and M. Reid, Fano 3-folds, K3 surfaces and graded rings, in Singapore Internat. Symp. in Topology and Geometry (NUS, 2001), Ed. A.J. Berrick and others, Contemp. Math. AMS, 2002; preprint math.AG/0202092, 29.

[BO] Alexei Bondal and Dmitri Orlov, Reconstruction of a variety from the derived category and groups of autoequivalences, Comp. Math. **125** (2001) 327–344.

[BKR] Tom Bridgeland, Alastair King and Miles Reid, Mukai implies McKay, J. Amer. Math. Soc. **14** (2001) 535–554.

[Ca] F. Campana, Connexité rationnelle des variétés de Fano, Ann. Sci. École Norm. Sup. **25** (1992) 539–545.

[Co] Alessio Corti, Factoring birational maps of threefolds after Sarkisov, J. Alg. Geom. **4** (1995) 223–254.

[Co2] Alessio Corti, Singularities of linear systems and 3-fold birational geometry, in [CR], 259–312.

[CM] Alessio Corti and Massimiliano Mella, Birational geometry of terminal quartic threefolds. I, can.dpmms.cam.ac.uk/∼corti/cm.pdf, 41.

[CPR] Alessio Corti, Sasha Pukhlikov and Miles Reid, Fano 3-fold hypersurfaces, in [CR], 175–258.

[CR] Alessio Corti and Miles Reid, Explicit birational geometry of 3-folds, CUP, 2000.

[CR2] Alessio Corti and Miles Reid, Weighted Grassmannians, in memorial volume for Paolo Francia (Genova, Sep 2001), M. Beltrametti Ed., de Gruyter 2002, 22, preprint math.AG/0206011, 27.

[Cr] Alastair Craw, An explicit construction of the McKay correspondence for A-Hilb $\mathbb{C}^3$, math.AG/0010053, 30.

[GHS] Tom Graber, Joe Harris and Jason Starr, Families of rationally connected varieties, math.AG/0109220, 21.

[Gr] Mark Gross, Examples of special Lagrangian fibrations, math.AG/0012002, 29.

[K] Kawamata Yujiro, Francia's flip and derived categories math.AG/0111041,

23.
[Ka1] Kawakita Masayuki, Divisorial contractions in dimension three which contract divisors to smooth points, Invent. Math. **145** (2001) 105–119.
[Ka2] Kawakita Masayuki, Divisorial contractions in dimension three which contract divisors to compound A_1 points, Comp. Math. to appear, math.AG/0010207, 23.
[Ko] Kollár János, Rational curves on algebraic varieties, Springer 1996.
[FA] Kollár János and others, Flips and abundance for algebraic threefolds, Astérisque **211** (1992), SMF 1992.
[KM] Kollár János and Mori Shigefumi, Birational geometry of algebraic varieties, CUP 1998.
[KMM] Kollár János, Miyaoka Yoichi and Mori Shigefumi, Rational connectedness and boundedness of Fano manifolds, J. Diff. Geom. **36** (1992) 765–779.
[KS] Maximilian Kreuzer and Harald Skarke, Calabi-Yau data, website tph16.tuwien.ac.at/∼kreuzer/CY.
[M] Matsuki Kenji, Introduction to the Mori program, Springer, 2002.
[O] Dmitri Orlov, Projective bundles, monoidal transformations, and derived categories of coherent sheaves, Izv. **56** (1992) 852–862 = Russian Acad. Sci. Izv. Math. **41** (1993) 133–141.
[PR] S. Papadakis and M. Reid, Kustin–Miller unprojection without complexes, to appear in J. Algebraic Geometry, math.AG/0011094, 18.
[YPG] Miles Reid, Young person's guide to canonical singularities, in Algebraic geometry (Bowdoin, 1985), Proc. Sympos. Pure Math. **46**, Part 1, AMS 1987, 345–414.
[Ki] Miles Reid, Graded rings and birational geometry, in Proc. of algebraic geometry symposium (Kinosaki, 2000), K. Ohno Ed., 1–72, get from www.maths.warwick.ac.uk/∼miles/3folds.
[Bou] Miles Reid, La correspondance de McKay, Séminaire Bourbaki, Astérisque **276**, SMF 2002, 53–72.
[Sh] V.V. Shokurov, Prelimiting flips, to appear in Proc. Steklov Inst., first draft Aug 1999, Mar 2002 draft 247, available from www.maths.warwick.ac.uk/∼miles/3folds.

ICM 2002 · Vol. II · 525–532

Sur les Algèbres Vertex Attachées aux Variétés Algébriques

Vadim Schechtman*

Abstract

Sganarelle: ... Mais encore faut-il croire quelque chose dans le monde: qu'est-ce donc que vous croyez?
Dom Juan: Ce que je crois?
Sganarelle: Oui.
Dom Juan: Je crois que deux et deux sont quatre, Sganarelle, et que quatre et quatre sont huit.
Molière, Dom Juan
One dicusses sheaves of vertex algebras over smooth varieties and their connections with characteristic classes.

2000 Mathematics Subject Classification: 17B69, 57R20.
Keywords and Phrases: Vertex algebras, Characteristic classes.

1. Introduction

Le but de cette note est de presenter une classification de certaines algèbres vertex, qui peuvent être associées à des variétés algébriques lisses; ceci est l'occasion de rencontrer des classes caractéristiques "style Pontryagin-Atiyah-Chern-Simons". Ceci a été obtenu dans [GMS] dont la presente note est un complément. On propose ici une définition plus simple d'une *algébroïde vertex* (*infra*, 2.4, 2.6), un énoncé plus précis et une démonstration courte du résultat principal de *op. cit.* (*infra*, 3.4, 3.6, 3.7). À la fin on propose une construction directe des algèbres vertex associées aux courbes, à l'aide des algèbres de Virasoro introduites par A.Beilinson. Le point de départ de cette note était une tentative de comprendre *le complexe de de Rham chiral* découvert par F.Malikov, [MSV]. Je remercie vivement mes amis et collaborateurs Arkady Vaintrob, Fyodor Malikov et Vassily Gorbounov.

*Laboratoire Emile Picard, Université Paul Sabatier, 118, Route de Narbonne, 31062 Toulouse Cedex 4, France. E-mail: schechtman@picard.ups-tlse.fr

2. Algébroïdes vertex

2.1. On fixe un corps de base k de caractéristique 0. Rappelons qu'une *algèbre vertex* est un k-espace vectoriel V muni d'un vecteur distingué $\mathbf{1} \in V$ (dit *vacuum*) et d'une famille d'applications k-linéaires $_{(i)} : V \otimes_k V \longrightarrow V$ $(i \in \mathbb{Z})$ telles que $x_{(i)}y = 0$ pour i assez grand. Si l'on pose $\partial x := x_{(-2)}\mathbf{1}$, on obtient un opérateur $\partial : V \longrightarrow V$. Les axiomes de [B] doivent être vérifiés. On n'est interessé que par les algèbres vertex $\mathbb{Z}_{\geq 0}$- *graduées*, ce qui signifie que l'espace V est muni d'une $\mathbb{Z}_{\geq 0}$-graduation (dite *poids conforme*), $V = \oplus_{n \geq 0} V_n$, $\mathbf{1} \in V_0$ et $V_{n(i)}V_m \subset V_{n+m-i-1}$. En particulier $\partial V_n \subset V_{n+1}$. Les morphismes de telles algèbres étant définis de manière évidente, on obtient la catégorie $\mathcal{V}ert$ des algèbres vertex $\mathbb{Z}_{\geq 0}$-graduées.

2.2. "Données classiques" associées à une algèbre vertex. Soit $V \in \mathcal{V}ert$. Pour être bref on écrira xy au lieu de $x_{(-1)}y$; c'est une opération non commutative et non associative en général. On a $V_nV_m \subset V_{n+m}$. Posons $A(V) = V_0$; l'opération xy est commutative et associative sur V_0; donc $A(V)$ devient une k-algèbre commutative avec unité $\mathbf{1}$. Posons $\mathcal{A}(V) = V_1$. Soit $\Omega(V)$ le sous-k-vectoriel de $\mathcal{A}(V)$ engendré par les éléments $a\partial b$, $a, b \in A(V)$. Alors $\Omega(V)$ devient un $A(V)$-module et $\partial : A(V) \longrightarrow \Omega(V)$ est une dérivation. En outre, si l'on pose $T(V) := \mathcal{A}(V)/\Omega(V)$, l'opération ax induit une structure de $A(V)$-module sur $T(V)$. (Par contre, $\mathcal{A}(V)$ n'est pas un $A(V)$-module en général, à cause de la non associativité de l'opération ax.) L'opération $_{(0)} : \mathcal{A}(V) \times \mathcal{A}(V) \longrightarrow \mathcal{A}(V)$ induit l'application $[,] : T(A) \times T(A) \longrightarrow T(A)$ qui est un crochet de Lie; l'opération $_{(0)} : \mathcal{A}(V) \times A(V) \longrightarrow A(V)$ induit une action de $T(V)$ sur $A(V)$ par dérivations; on a $[\tau, a\tau'] = a[\tau, \tau'] + \tau(a)[\tau, \tau']$, i.e. $T(V)$ devient une $A(V)$-*algébroïde de Lie*. La première opération induit aussi une action de l'algèbre de Lie $T(V)$ sur $\Omega(V)$ telle que ∂ est un morphisme de $T(V)$-modules, et $\tau(a\omega) = \tau(a)\omega + a\tau(\omega)$. Enfin, l'opération $_{(1)} : \mathcal{A}(V) \times \mathcal{A}(V) \longrightarrow A$ est symétrique et induit un accouplement $A(V)$-bilinéaire $\langle,\rangle : T(V) \times \Omega(V) \longrightarrow A(V)$ telle que $\tau(\langle\tau', \omega\rangle) = \langle[\tau, \tau'], \omega\rangle + \langle\tau', \tau(\omega)\rangle$ et $(a\tau)(\omega) = a\tau(\omega) + \langle\tau, \omega\rangle\partial a$.

2.3. "Données quantiques." Les propriétés (Alg1) — (Alg3) ci-dessous sont vérifiées, où $a \in A(V)$, $x, y, z \in \mathcal{A}(V)$, $\pi : A(V) \longrightarrow T(A)$ étant la projection canonique.

(Alg1) $(ax)_{(1)}y = a(x_{(1)}y) - \pi(x)\pi(y)(a)$.

(Alg2) $x_{(0)}y + y_{(0)}x = \partial(x_{(1)}y)$; $(\partial x)_{(0)}y = 0$.

(Alg3) $x_{(0)}(y_{(i)}z) = (x_{(0)}y)_{(i)}z + y_{(i)}(x_{(0)}z)$, $i = 0, 1$.

2.4. Soit A une k-algèbre commutative de type fini, lisse sur k. Posons $\Omega(A) = \Omega^1_{A/k}$, $T(A) = Der_k(A, A)$ (l'algèbre de Lie de k-dérivations de A). Soit $\partial = \partial_{DR} : A \longrightarrow \Omega(A)$ la dérivation universelle. On a l'accouplement non degenéré A-bilinéaire $\langle,\rangle : T(A) \times \Omega(A) \longrightarrow A$; l'algèbre de Lie $T(A)$ agit sur $\Omega(A)$ par la dérivée de Lie. Ces données vérifient toutes les propriétés de 2.2.

Une A-algébroïde vertex est un k-espace vectoriel $\mathcal{A}$ muni d'un sous-espace $F^1\mathcal{A} \subset \mathcal{A}$ avec les identifications de k-vectoriels $F^1\mathcal{A} = \Omega(A)$, $\mathcal{A}/F^1\mathcal{A} = T(\mathcal{A})$ et des opérations k-bilinéaires $_{(-1)} : A \times \mathcal{A} \longrightarrow \mathcal{A}$, $(a, x) \mapsto ax = a_{(-1)}x$, $_{(1)} :$

$\mathcal{A} \times \mathcal{A} \longrightarrow A$ symétrique, $_{(0)} : \mathcal{A} \times \mathcal{A} \longrightarrow \mathcal{A}$. On demande que (i) $A_{(-1)}\Omega(A) \subset \Omega(A)$ et que l'action de A sur $\Omega(A)$ et sur $T(A)$ induite par $_{(-1)}$ coïncide avec l'action canonique; (ii) $\Omega(A)_{(i)}\Omega(A) = 0$ $(i = 0, 1)$; $\Omega_{(0)}\mathcal{A} \subset \Omega(A)$, l'opération $T(A) \times T(A) \longrightarrow T(A)$ induite par $_{(0)}$ coïncide avec le crochet de Lie, et l'action induite $T(A) \times \Omega(A) \longrightarrow \Omega(A)$ coïncide avec la derivée de Lie; (iii) l'acccouplement $\langle , \rangle : T(A) \times \Omega(A) \longrightarrow A$ induit par $_{(1)}$ coïncide avec l'accouplement canonique. Enfin, les propriétés (Alg1) — (Alg3) doivent être vérifiées. Dans (Alg3) pour $i = 1$ on interprète la partie de gauche comme $\pi(x)(y_{(1)}z)$.

Soit $T^{\cdot}(A)$ une algébre de Lie dg concentrée en degrés $-1, 0$, avec $T^{-1}(A) = T^0(A) = T(A)$, $d : T^{-1}(A) \longrightarrow T^0(A)$ l'identité, le crochet $[,]_{0,-1}$ l'action adjointe. Soit $\Omega^{\cdot}(A) : 0 \longrightarrow A \xrightarrow{\partial} \Omega(A) \longrightarrow \Omega(A)/\partial A \longrightarrow 0$ le complexe concentré en degrés $-2, -1, 0$ avec les différentielles évidentes. Ce complexe est un module dg sur $T^{\cdot}(A)$ (l'action de $T^0(A)$ est par la dérivée de Lie, la composante $[,]_{-1,-1} : T^{-1}(A) \times \Omega^{-1}(A) \longrightarrow \Omega^{-2}(A)$ étant l'accouplement canonique, et la composante $[,]_{-1,0}$ étant définie par $[\tau, \bar{\omega}] = i_\tau(d\omega)$, où $\bar{\omega} \in \Omega(A)/\partial A$ est l'image de $\omega \in \Omega(A)$, $d : \Omega^1_{A/k} \longrightarrow \Omega^2_{A/k}$ est la différentielle de de Rham, $i_\tau : \Omega^2_{A/k} \longrightarrow \Omega^1_{A/k}$ est la convolution avec τ). On peut exprimer les axiomes (Alg2) et (Alg3) en disant que l'on a une algèbre de Lie dg $\mathcal{A}^{\cdot} : 0 \longrightarrow A \longrightarrow \mathcal{A} \longrightarrow \mathcal{A}/\partial A \longrightarrow 0$, concentrée en degrés $-2, -1, 0$, extension de $T^{\cdot}(A)$ par $\Omega^{\cdot}(A)$ (considérée comme une sous-algèbre de Lie abélienne), telle que l'action de $T^{\cdot}(A)$ sur $\Omega^{\cdot}(A)$ induite coïncide avec celle décrite ci-dessus. Un morphisme $g : \mathcal{A} \longrightarrow \mathcal{A}'$ est une application k-linéaire respectant les opérations $_{(i)}$ et les filtrations, qui induit l'identité sur $\Omega(A), T(A)$. D'où la catégorie $\mathcal{A}lg_A$ des A-algébroïdes vertex, qui est *un groupoïde* (chaque morphisme est un isomorphisme).

2.5. Soit A comme dans 2.4. On définit la catégorie $\mathcal{V}ert_A$ dont les objets sont $V \in \mathcal{V}ert$ munies d'un isomorphisme de k-algèbres $A(V) \xrightarrow{\sim} A$, cet isomorphisme identifiant les données classiques $(T(V), \Omega(V), \partial, \langle , \rangle)$ correspondantes avec les données standardes $(T(A), \Omega(A), \partial_{DR}, \langle , \rangle)$ décrites dans 2.4. Les morphismes sont les morphismes des algèbres vertex induisants l'identité sur les données classiques.

La construction 2.2, 2.3 donne lieu au foncteur $\mathcal{A}lg : \mathcal{V}ert_A \longrightarrow \mathcal{A}lg_A$, $V \mapsto \mathcal{A}(V)$. Ce foncteur admet l'adjoint à gauche $U : \mathcal{A}lg_A \longrightarrow \mathcal{V}ert_A$, l'algèbre vertex $U\mathcal{A}$ étant appelée *l'algèbre enveloppante* d'un algébroïde vertex $\mathcal{A}$. Pour chaque $\mathcal{A} \in \mathcal{A}lg_A$ le morphisme d'adjonction $\mathcal{A} \longrightarrow \mathcal{A}lg(U\mathcal{A})$ est un isomorphisme.

2.6. Le langage suivant est un peu plus explicite et est parfois commode. Appelons *A-algébroïde vertex scindée* un couple $\mathcal{B} = (\langle , \rangle, c)$, où $\langle , \rangle : T(A) \times T(A) \longrightarrow A$ (resp. $c : T(A) \times T(A) \longrightarrow \Omega(A)$) est une application k-bilinéaire symétrique (resp. antisymétrique). On demande que les propriétés (AlgScind1)–(AlgScind3) ci-dessous soient vérifiées.

(AlgScind1) $\langle a\tau, b\tau' \rangle - a\langle \tau, b\tau' \rangle - b\langle a\tau, \tau' \rangle + ab\langle \tau, \tau' \rangle = -\tau'(a)\tau(b)$.

(AlgScind2) $\langle \tau'', c(\tau, \tau') \rangle + \langle \tau', c(\tau, \tau'') \rangle = \langle [\tau, \tau'], \tau'' \rangle + \langle \tau', [\tau, \tau''] \rangle - \tau(\langle \tau', \tau'' \rangle) + \tau'(\langle \tau, \tau'' \rangle)/2 + \tau''(\langle \tau, \tau' \rangle)/2$.

(AlgScind3) $3\{\tau(c(\tau',\tau'')) + \tau'(c(\tau'',\tau)) + \tau''(c(\tau,\tau')) - c([\tau,\tau'],\tau'') - c([\tau',\tau''],\tau) - c([\tau'',\tau],\tau')\} = \partial\{\langle\tau,\frac{1}{2}[\tau',\tau''] + c(\tau',\tau'')\rangle + \langle\tau',\frac{1}{2}[\tau'',\tau] + c(\tau'',\tau)\rangle + \langle\tau'',\frac{1}{2}[\tau,\tau'] + c(\tau,\tau')\rangle\}$.

Ses propriétés entraînent que $\langle,\rangle$ et c sont des opérateurs différentiels, d'ordres 2 et 3 respectivement.

Étant donnés deux A-algébroïdes vertex scindées $\mathcal{B} = (\langle,\rangle, c)$ et $\mathcal{B}' = (\langle,\rangle', c')$, *un morphisme* $f : \mathcal{B} \longrightarrow \mathcal{B}'$ est par définition une application k-linéaire $h = h_f : T(A) \longrightarrow \Omega(A)$ satisfaisant les propriétés (Mor1)–(Mor3) ci-dessous (dont la première implique que h est un opérateur différentiel d'ordre 2).

(Mor1) $\langle\tau', h(a\tau)\rangle - \langle a\tau,\tau'\rangle + \langle a\tau,\tau'\rangle' = a\{\langle\tau', h(\tau)\rangle - \langle\tau,\tau'\rangle + \langle\tau',\tau\rangle'\}$.

(Mor2) $\langle\tau,\tau'\rangle - \langle\tau,\tau'\rangle' = \langle\tau, h(\tau')\rangle + \langle\tau', h(\tau)\rangle$.

(Mor3)

$c(\tau,\tau') - c'(\tau,\tau') = \tau'(h(\tau)) - \tau(h(\tau')) + h([\tau,\tau']) + \partial\{\langle\tau, h(\tau')\rangle - \langle\tau', h(\tau)\rangle\}/2$.

La composition est définie par $h_{ff'} = h_f + h_{f'}$; l'identité est $h_{id} = 0$. D'où on obtient le groupoïde $\mathcal{AlgScind}_A$ des A-algébroïdes vertex scindées. Étant donné $\mathcal{B}$ comme ci-dessus, on pose $\mathcal{A}(\mathcal{B}) = T(A) \oplus \Omega(A)$ et définit les opérations $_{(i)}$, $i = -1, 0, 1$ par les formules $a_{(-1)}\tau = (a\tau, -\gamma(a,\tau))$, où $\gamma(a,\tau) \in \Omega(A)$ est défini par

$$\langle\tau', \gamma(a,\tau)\rangle = \langle a\tau,\tau'\rangle - a\langle\tau',\tau\rangle + \tau\tau'(a)$$

(l'axiome (AlgScind1) signifie que cette expression est A-linéaire en τ'); $\tau_{(0)}\tau' = ([\tau,\tau'], -c(\tau,\tau') + \frac{1}{2}\partial\langle\tau,\tau'\rangle$, $\tau_{(1)}\tau' = \langle\tau,\tau'\rangle$. Ceci définit $\mathcal{A}(\mathcal{B}) \in \mathcal{Alg}_A$. Si $f : \mathcal{B} \longrightarrow \mathcal{B}'$ est comme ci-dessus, on définit le morphisme $g(f) : \mathcal{A}(\mathcal{B}) \longrightarrow \mathcal{A}(\mathcal{B}')$ par $g(f)(\tau) = (\tau, h_f(\tau))$. Ceci définit un foncteur $\mathcal{AlgScind}_A \longrightarrow \mathcal{Alg}_A$ qui est une équivalence des catégories.

2.7. Exemple. Supposons que A est telle que $T(A)$ soit un A-module libre et il existe une A-base $\mathfrak{b} = \{\tau_1, \ldots, \tau_n\}$ de $T(A)$ telle que $[\tau_i, \tau_j] = 0$ pour tous i, j. Nous appelerons telles algèbres *petites* est les bases $\mathfrak{b}$ *abéliennes.* On pose

$$\langle a\tau_i, b\tau_j\rangle_{\mathfrak{b}} = -b\tau_i\tau_j(a) - a\tau_j\tau_i(b) - \tau_i(b)\tau_j(a), \qquad (2.7)_{\langle,\rangle}$$

$$c_{\mathfrak{b}}(a\tau_i, b\tau_j) = \frac{1}{2}\{\tau_i(b)\partial\tau_j(a) - \tau_j(a)\partial\tau_i(b)\} + \frac{1}{2}\partial\{b\tau_i\tau_j(a) - a\tau_j\tau_i(b)\}. \qquad (2.7)_c$$

Alors $\mathcal{B}_{\mathfrak{b}} = (\langle,\rangle_{\mathfrak{b}}, c_{\mathfrak{b}})$ est une A-algébroïde vertex scindée.

3. Classification

3.1. Une algèbre A étant toujours comme dans 2.4, on définit un groupoïde $\mathcal{G}r(\Omega_A^{[2,3\rangle})$ dont les objets sont les formes différentielles *fermées* $\omega \in \Omega^{3,fer}_{A/k}$, avec

$$Hom_{\mathcal{G}r(\Omega_A^{[2,3\rangle})}(\omega,\omega') = \{\eta \in \Omega^2_{A/k} | d\eta = \omega - \omega'\}.$$

La composition des morphismes est l'addition de 2-formes. L'addition des 3-formes induit une structure d'un *groupe abélien en catégories* sur ce groupoïde.

On remarque que si $\mathcal{A}, \mathcal{A}'$ sont deux A-algébroïdes vertex avec le même espace sous-jacent la même opération $_{(1)}$, alors $x_{(0)}y - x_{(0)'}y \in \Omega(A)$; cet élément ne dépend que des $\pi(x), \pi(y)$, où $\pi : \mathcal{A} \longrightarrow T(A)$ est l'application canonique, d'où l'application $c_{\mathcal{A},\mathcal{A}'} : T(A) \times T(A) \longrightarrow \Omega(A)$. De plus, cette application est A-bilinéaire, et $\omega_{\mathcal{A},\mathcal{A}'}(\tau, \tau', \tau'') := \langle \tau, c_{\mathcal{A},\mathcal{A}'}(\tau', \tau'')\rangle$ est antisymétrique en τ, τ', τ'', donc $\omega_{\mathcal{A},\mathcal{A}'}$ peut être considerée comme une 3-forme différentielle, et cette forme est fermée.

Réciproquement, étant donné $\mathcal{A} = \mathcal{A}lg_A$ et $\omega \in \Omega^{3,fer}_{A/k}$, on définit $\mathcal{A}' = \mathcal{A} \dot{+} \omega \in \mathcal{A}lg_A$ ayant le même espace sous-jacent que $\mathcal{A}$ et la même opération $_{(1)}$, avec $_{(0)'} =_{(0)} -\omega$.

Si $g : \mathcal{A} \dot{+} \omega \longrightarrow \mathcal{A} \dot{+} \omega'$ est un morphisme, alors $(g - id)(\mathcal{A}) \subset \Omega(A)$, $(g - id)|_{\Omega(A)} = 0$, donc $g - id$ induit une application $h_g : T(A) \longrightarrow \Omega(A)$. La fonction, $\eta_g(\tau, \tau') := \langle \tau, h_g(\tau)\rangle$ est antisymétrique en τ, τ' et A-bilinéaire, donc peut être considerée comme une 2-forme différentielle; on a $d\eta = \omega - \omega'$. Ceci induit une bijection $Hom_{\mathcal{A}lg_A}(\mathcal{A} \dot{+} \omega, \mathcal{A} \dot{+} \omega') = \{\eta \in \Omega^2_{A/k} | d\eta = \omega - \omega'\}$. On a $Hom_{\mathcal{A}lg_A}(\mathcal{A}, \mathcal{A}') = Hom_{\mathcal{A}lg_A}(\mathcal{A} \dot{+} \omega, \mathcal{A}' \dot{+} \omega)$. Cela définit une Action

$$\dot{+}: \mathcal{A}lg_A \times \mathcal{G}r(\Omega_A^{[2,3\rangle}) \longrightarrow \mathcal{A}lg_A. \tag{3.1.1}$$

3.2. Théorème. *Si A est petite* (voir 2.7), *alors le groupoïde $\mathcal{A}lg_A$ est un Torseur sous $\mathcal{G}r(\Omega_A^{[2,3\rangle})$ par rapport à l'Action* (3.1.1), *c'est à dire, pour chaque $\mathcal{A} \in \mathcal{A}lg_A$ le foncteur $\mathcal{A} \dot{+}? : \mathcal{G}r(\Omega_A^{[2,3\rangle}) \longrightarrow \mathcal{A}lg_A$ est une équivalence.*

Par exemple, l'ensemble $\pi_0(\mathcal{A}lg_A)$ des classes d'isomorphisme de $\mathcal{A}lg_A$ est un torseur sous $H^3_{DR}(A)$. Grace à 2.6 le Torseur $\mathcal{A}lg_A$ est non-vide pour A petite.

3.3. Soient A petite, et $\mathfrak{b} = \{\tau_i\}, \mathfrak{b}' = \{\tau'_i\}$ deux bases abéliennes, d'où les algébroïdes scindées $\mathcal{B}_\mathfrak{b}, \mathcal{B}_{\mathfrak{b}'}$; on a $\tau'_i = \phi^{ij}\tau_j$ (la règle de Einstein est sous-entendue), $\phi = (\phi^{ij}) \in GL_n(A)$ (pour être bref, on écrit $\mathfrak{b}' = \phi\mathfrak{b}$). On définit une application $h_{\mathfrak{b}',\mathfrak{b}} : T(A) \longrightarrow \Omega(A)$, comme étant l'unique opérateur satisfaisant (Mor1), tel que $\langle \tau'_i, h_{\mathfrak{b}',\mathfrak{b}}(\tau'_j)\rangle = -\frac{1}{2}\langle \tau'_i, \tau'_j\rangle_\mathfrak{b}$. De plus, on définit une application $c_{\mathfrak{b}',\mathfrak{b}} : T(A) \times T(A) \longrightarrow \Omega(A)$ comme étant l'unique opérateur tel que $\mathcal{B}_{\mathfrak{b}',\mathfrak{b}} := (\langle,\rangle_\mathfrak{b}, c_{\mathfrak{b}',\mathfrak{b}})$ soit une algébroïde vertex scindée, et $h_{\mathfrak{b}',\mathfrak{b}}$ soit un morphisme d'algébroïdes scindées $\mathcal{B}_{\mathfrak{b}'} \longrightarrow \mathcal{B}_{\mathfrak{b}',\mathfrak{b}}$. D'où la 3-forme $\alpha_{\mathfrak{b}',\mathfrak{b}} \in \Omega^{3,fer}_{A/k}$ telle que $\mathcal{B}_\mathfrak{b} = \mathcal{B}_{\mathfrak{b}',\mathfrak{b}} \dot{+} \alpha_{\mathfrak{b}',\mathfrak{b}}$. Si $\mathfrak{b}'' = \{\tau''_i\}$ est la troisième base abélienne, avec $\tau''_i = \psi^{ij}\tau'_j$, on définit la 2-forme $\beta_{\mathfrak{b}'',\mathfrak{b}',\mathfrak{b}} := h_{\mathfrak{b}'',\mathfrak{b}'} + h_{\mathfrak{b}',\mathfrak{b}} - h_{\mathfrak{b}'',\mathfrak{b}} \in \Omega^2_{A/k}$.

3.4. Théorème. $\beta_{\mathfrak{b}'',\mathfrak{b}',\mathfrak{b}} = \frac{1}{2}tr\{\phi^{-1}\psi^{-1}d\psi d\phi\}$, $\alpha_{\mathfrak{b}',\mathfrak{b}} = \frac{1}{6}tr\{(\phi^{-1}d\phi)^3\}$.

Démonstration. Il resulte de (2.7) que

$$c_\mathfrak{b}(\tau'_i, \tau'_j) = \frac{1}{2}tr\big\{\phi^{-1}\tau'_j(\phi)\phi^{-1}\partial\tau'_i(\phi) - \phi^{-1}\tau'_i(\phi)\phi^{-1}\tau'_j(\phi)\phi^{-1}\partial\phi - (i \leftrightarrow j)\big\}, \tag{3.4}_c$$

$$\langle \tau'_i, \tau'_j\rangle_\mathfrak{b} = tr\big\{-2\phi^{-1}\tau'_i\tau'_j(\phi) + \phi^{-1}\tau'_i(\phi)\phi^{-1}\tau'_j(\phi)\big\} \tag{3.4}_{\langle,\rangle}$$

d'où, en utilisant (Mor1),

$$h_{\mathfrak{b}',\mathfrak{b}}(\tau''_i) = tr\big\{\phi^{-1}\partial\tau''_i(\phi) - \frac{1}{2}\phi^{-1}\tau''_i(\phi)\phi^{-1}\partial\phi + \phi^{-1}\psi^{-1}\tau''_i(\psi)\partial\phi\big\}. \tag{3.4}_h$$

Par définition, $c_{\mathfrak{b}',\mathfrak{b}}(\tau_i',\tau_j') = \tau_i'(h(\tau_j')) - \tau_j'(h(\tau_i'))$; $\alpha_{\mathfrak{b}',\mathfrak{b}} = c_{\mathfrak{b}} - c_{\mathfrak{b}',\mathfrak{b}}$, d'où;

$$\alpha_{\mathfrak{b}',\mathfrak{b}}(\tau_i',\tau_j') = -\frac{1}{2}tr\{\phi^{-1}\tau_i'(\phi)\phi^{-1}\tau_j'(\phi)\phi^{-1}\partial\phi - (i \leftrightarrow j)\}. \qquad (3.4)_\alpha$$

En outre, $(3.4)_h$ entraîne

$$\beta_{\mathfrak{b}'',\mathfrak{b}',\mathfrak{b}}(\tau_i'') = \frac{1}{2}tr\{\phi^{-1}\psi^{-1}\tau_i''(\psi)\partial\phi - \tau_i''(\phi)\phi^{-1}\psi^{-1}\partial\psi\} \qquad (3.4)_\beta$$

d'où le théorème. Ici l'on identifie une 3-forme α avec une application antisymétrique $T(A)\times T(A) \longrightarrow \Omega(A)$ définie par $\alpha(\tau,\tau') = i_\tau i_{\tau'}\alpha$. $\triangle$

3.5. Classe de Pontryagin. Soit X une variété algébrique lisse sur k, E un fibré vectoriel sur X. Choisissons une recouvrement affine $\mathfrak{U} = \{U_i\}$ de X, et des bases $\mathfrak{b}^i$ des $\Gamma(U_i,\mathcal{O}_X)$-modules $\Gamma(U_i,E)$, d'où le cocycle de $\overset{\vee}{C}$ech $\phi = (\phi_{ij})$, $\phi_{ij} \in \Gamma(U_{ij},GL_n(\mathcal{O}_X))$, $\mathfrak{b}^i = \phi_{ij}\mathfrak{b}^j$ sur U_{ij}, $\phi_{ij}\phi_{jk} = \phi_{ik}$ sur U_{ijk}. Considérons les cochaînes de $\overset{\vee}{C}$ech $p_2(\phi) = \big(\frac{1}{2}tr\{\phi_{jk}^{-1}\phi_{ij}^{-1}d\phi_{ij}d\phi_{jk}\}\big) \in C^2(\mathfrak{U},\Omega_X^2)$, $p_3(\phi) = \big(\frac{1}{6}tr\{(\phi_{ij}^{-1}d\phi_{ij})^3\}\big) \in C^1(\mathfrak{U},\Omega_X^3)$; on a $d_{\overset{\vee}{C}ech}p_2(\phi) = 0$, $d_{DR}p_2(\phi) = d_{\overset{\vee}{C}ech}p_3(\phi), d_{DR}p_3(\phi) = 0$. Il en résulte que $p(\phi) := \big(p_2(\phi),p_3(\phi)\big) \in Z^2(\mathfrak{U},\Omega_X^{[2,3\rangle})$ où $\Omega_X^{[2,3\rangle} := \big(\Omega_X^2 \longrightarrow \Omega_X^{3,fer}\big)$, la différentielle totale dans le bicomplexe de $\overset{\vee}{C}$ech à coefficients dans ce complexe étant $d = d_{DR} + (-1)^{|DR|}d_{\overset{\vee}{C}ech}$. De plus, si l'on choisit des autres bases ${}'\mathfrak{b}^i = g_i\mathfrak{b}^i$, d'ù $g = (g_i) \in C^0(\mathfrak{U},GL_n(\mathcal{O}_X))$, le cocycle correspondant est $\phi' = {}^g\phi$, où ${}^g\phi_{ij} = g_i\phi_{ij}g_j^{-1}$. On définit

$$p_2(\phi,g) := \big(\frac{1}{2}tr\{\phi_{ij}^{-1}g_i^{-1}dg_i\phi_{ij}g_j^{-1}dg_j + \phi_{ij}^{-1}d\phi_{ij}g_j^{-1}dg_j - g_i^{-1}dg_id\phi_{ij}\phi_{ij}^{-1}\}\big),$$

$$p_3(g) := \big(\frac{1}{6}tr\{(g_i^{-1}dg_i)^3\}\big);\ p(\phi,g) = \big(p_2(\phi,g),p_3(g)\big) \in C^1(\mathfrak{U},\Omega_X^{[2,3\rangle}).$$

Alors $p_3({}^g\phi) = p_3(\phi) + d_{\overset{\vee}{C}ech}p_3(g) + d_{DR}p_2(\phi,g), p_2(\phi^g) = p_2(\phi) + d_{\overset{\vee}{C}ech}p_2(\phi,g)$, d'où $p({}^g\phi) = p(\phi) + dp(\phi,g)$. Donc la classe $p(E)$ de $p(\phi)$ dans $H^2(X,\Omega_X^{[2,3\rangle})$ qu'on peut appeler *la classe de Pontryagin-Atyiah-Chern-Simons (pacs)*, ne depend que de E. On remarque que $p(\phi) = p(\phi^{-1t})$, donc $p(E) = p(E^*)$.

3.6. Les groupoïdes $\mathcal{A}lg_{\Gamma(U,\mathcal{O}_X)}$, $U \subset X$, forment un champ $\mathfrak{Alg}_X$ sur la topologie de Zariski (même étale), parsque les opérations ${}_{(i)}$ sont des opérateurs différentiels qui se localisent. D'après 3.2, $\mathfrak{A}lg_X$ est une *gerbe* sous $\Omega_X^{[2,3\rangle}$ (localement non-vide, mais pas localement connexe). Donc *la classe caractéristique* $c(\mathfrak{Alg}_X) \in H^2(X,\Omega_X^{[2,3\rangle})$ est définie, telle que $c(\mathfrak{Alg}_X) = 0$ ssi $\Gamma(X,\mathfrak{A}lg_X)$ est non-vide. Rappelons sa définition. On choisit un recouvrement affine $\mathfrak{U} = \{U_i\}$ de X avec U_i petites; on choisit les objets $\mathcal{A}_i \in \Gamma(U_i,\mathfrak{Alg}_X)$. Sur les doubles intersections, il existe les isomorphismes h_{ij} : $\mathcal{A}_j|_{U_{ij}} \xrightarrow{\sim} \mathcal{A}_i|_{U_{ij}} + \alpha_{ij}$, $\alpha_{ij} \in \Omega^{3,fer}(U_{ij})$. Si l'on pose $\beta_{ijk} := h_{ij}|_{U_{ijk}} - h_{ik}|_{U_{ijk}} + h_{jk}|_{U_{ijk}} \in \Omega^2(U_{ijk})$, on a $c(\{\mathcal{A}_i\},\{h_{ij}\}) :=$

$((\alpha_{ij}),(\beta_{ijk})) \in Z^2(\mathfrak{U},\Omega_X^{[2,3\rangle})$. Pour une autre famille $(\{\mathcal{A}'_i\},\{h'_{ij}\})$ il existent $h_i : \mathcal{A}_i \xrightarrow{\sim} \mathcal{A}'_i \dotplus \alpha_i$, $\alpha_i \in \Omega^{3,fer}_{A/k}$; alors $(h_j \dotplus \alpha_{ij}) \circ h_{ij} : \mathcal{A}_i \xrightarrow{\sim} \mathcal{A}'_j \dotplus (\alpha_j + \alpha_{ij})$ et $(h'_{ij} \dotplus \alpha_i) \circ h_i : \mathcal{A}_i \xrightarrow{\sim} \mathcal{A}'_j \dotplus (\alpha'_{ij} + \alpha_i)$, donc il existe l'unique $\beta_{ij} \in \Omega^2_{A/k}$ telle que $d\beta_{ij} = \alpha'_{ij} - \alpha_{ij} + \alpha_i - \alpha_j$. Alors $d((\alpha_i),(\beta_{ij})) = c(\{\mathcal{A}'_i\},\{h'_{ij}\}) - c(\{\mathcal{A}_i\},\{h_{ij}\})$; par définition $c(\mathfrak{Alg}_X)$ est la classe de $c(\{\mathcal{A}_i\},\{h_{ij}\})$ dans la cohomologie.

Soit $\mathcal{T}_X$ le fibré tangent de X. Choisissons des bases bonnes $\mathfrak{b}^i$ de $\Gamma(U_i,\mathcal{T}_X)$, avec $\mathfrak{b}^i = \phi_{ij}\mathfrak{b}^j$, $\phi = (\phi_{ij}) \in Z^1(\mathfrak{U},GL_n(\mathcal{O}_X))$. Alors, d'après 3.4, $\alpha_{\mathfrak{b}^i\mathfrak{b}^j} = p_3(\phi)_{ij}$ et $\beta_{\mathfrak{b}^i\mathfrak{b}^j\mathfrak{b}^l} := h_{\mathfrak{b}^j\mathfrak{b}^l} - h_{\mathfrak{b}^i\mathfrak{b}^l} + h_{\mathfrak{b}^i\mathfrak{b}^j} = p_2(\phi)_{ijl}$. De plus, si $\{'\mathfrak{b}^i\}$ est une autre famille des bases bonnes, avec ${}'\mathfrak{b}^i = g_i\mathfrak{b}^i$, $g = (g_i)$, alors $\alpha_{'\mathfrak{b}^i\mathfrak{b}^i} = p_3(g)_i$ et $h_{'\mathfrak{b}^i\,'\mathfrak{b}^j} - h_{\mathfrak{b}^i\mathfrak{b}^j} + h_{'\mathfrak{b}^i\mathfrak{b}^i} - h_{'\mathfrak{b}^j\mathfrak{b}^j} = p_2(\phi,g)_{ij}$. En particulier, on a

3.7. Théorème. *$c(\mathfrak{Alg}_X) = p(\mathcal{T}_X)$, où $\mathcal{T}_X$ est le fibré tangent de X.* $\triangle$

Soit ϕ comme dans 3.6, $p = p(\phi)$; soit $\mathcal{G}r_p$ le groupoïde dont les objets sont les 1-cochaînes de Cech $\omega \in C^1(\mathfrak{U},\Omega_X^{[2,3\rangle})$ telles que $d\omega = p$, avec $Hom_{\mathcal{G}r_p}(\omega,\omega') = \{\eta \in C^0(\mathfrak{U},\Omega_X^{[2,3\rangle}) |\ d\eta = \omega - \omega'\}$. La construction 3.6 donne lieu au foncteur $\mathcal{G}_p \longrightarrow \Gamma(X,\mathfrak{Alg}_X)$ qui est une équivalence des catégories. Il en résulte que $\pi_0\Gamma(X,\mathfrak{Alg}_X)$ est un torseur sous $H^1(X,\Omega_X^{[2,3\rangle})$, non-vide si et seulement si $p(\mathcal{T}_X) = 0$, et pour $\mathcal{A} \in \Gamma(X,\mathfrak{Alg}_X)$ le groupe $Aut(\mathcal{A})$ est isomorphe à $H^0(X,\Omega_X^{[2,3\rangle})$.

4. Exemple

Soit X une courbe lisse sur k. Dans ce cas $\Omega_X^{[2,3\rangle} = 0$, donc sur X il existe l'unique, à isomorphisme unique près, $\mathcal{O}_X$-algébroïde vertex $\mathcal{A}_X$. On propose ici une construction directe de $\mathcal{A}_X$. Pour $j \in \mathbb{Z}$ on a défini dans [BS] le faisceau d'algèbres de Lie différentielles graduées $\mathcal{A}^\cdot_j$ (*j-ième Virasoro*) sur X (cf. [BS] 3.1). On a $\mathcal{A}^i_j = 0$ pour $i \neq -2,-1,0$; $\mathcal{A}^{-2}_j = \mathcal{O}_X$, $\mathcal{A}^0_j = \mathcal{T}_X$. On a la suite exacte canonique des k-vectoriels (des $\mathcal{O}_X$-modules si $j = 0$) $0 \longrightarrow \Omega^1_X \longrightarrow \mathcal{A}^{-1}_j \xrightarrow{\pi} \mathcal{T}_X \longrightarrow 0$; Par définition, la différentielle $d : \mathcal{A}^{-1}_j \longrightarrow \mathcal{A}^0_j$ est égale à π et $d : \mathcal{A}^{-2}_j \longrightarrow \mathcal{A}^{-1}_j$ est égale à la composée de la différentielle de de Rham avec l'inclusion $\Omega^1_X \hookrightarrow \mathcal{A}^{-1}_j$. Comme il est expliqué dans *op. cit.*, la catégorie des algèbres de Lie dg comme ci-dessus est un *k-espace vectoriel en catégories*; en particulier, on peut les multiplier par un scalaire. On a l'isomorphisme canonique $\mathcal{A}^\cdot_j \xrightarrow{\sim} (6j^2 - 6j + 1)\mathcal{A}^\cdot_0$. Pour chaque $\lambda \in k$ on a l'isomorphisme des k-modules $\lambda\mathcal{A}^{-1}_0 \xrightarrow{\sim} \mathcal{A}^{-1}_0$, donc la structure canonique d'un $\mathcal{O}_X$-module sur $\mathcal{A}^{-1}_0$ induit une structure de $\mathcal{O}_X$-module sur $\lambda\mathcal{A}^{-1}_0$.

Considérons l'algèbre de Lie dg $6\mathcal{A}_0$. On pose $\mathcal{A}_X = 6\mathcal{A}^{-1}_0$. On définit les opérations par $a_{(-1)}x = ax - 2\partial\pi(x)(a)$; $x_{(0)}y = [\pi(x),y]$, $x_{(1)}y = [x,y]$ ($a \in \mathcal{O}_X$, $x,y \in \mathcal{A}_X$). Alors les axiomes (Alg1)–(Alg3) sont vérifiés, et l'on obtient une algébroïde vertex sur X.

Bibliographie

[BS] A. Beilinson, V. Schechtman, Determinant bundles and Virasoro algebras, *Comm. Math. Phys.*, 118 (1988), 651–701.

[B] R. Borcherds, Vertex algebras, Kac-Moody algebras, and the Monster, *Proc. Natl. Acad. Sci, USA*, 83 (May 1986), 3068–3071.

[GMS] V. Gorbounov, F. Malikov & V. Schechtman, *Gerbes of chiral differential operators. II*, math.AG/0003170.

[MSV] F. Malikov, V. Schechtman & A. Vaintrob, Chiral de Rham complex, *Comm. Math. Phys.*, 204 (1999), 439–473.

ICM 2002 · Vol. II · 533–541

Topology of Singular Algebraic Varieties

B. Totaro*

Abstract

I will discuss recent progress by many people in the program of extending natural topological invariants from manifolds to singular spaces. Intersection homology theory and mixed Hodge theory are model examples of such invariants. The past 20 years have seen a series of new invariants, partly inspired by string theory, such as motivic integration and the elliptic genus of a singular variety. These theories are not defined in a topological way, but there are intriguing hints of their topological significance.

2000 Mathematics Subject Classification: 14F43, 32S35, 58J26.
Keywords and Phrases: Intersection homology, Weight filtration, Elliptic genus.

1. Introduction

The most useful fact about singular complex algebraic varieties is Hironaka's theorem that there is always a resolution of singularities [20]. It has long been clear that the non-uniqueness of resolutions poses a difficulty in many applications. Many different methods have been used to get around this difficulty so as to define invariants of singular varieties. One approach is to try to describe the relation between any two resolutions, leading to ideas such as cubical hyperresolutions [18] and the weak factorization theorem ([1], [31]). Another idea, coming from minimal model theory, is to insist on the special importance of crepant resolutions, and more generally to emphasize the role of the canonical bundle. Recently the interplay between these two approaches has been very successful, as I will describe.

The recent methods tend to be more roundabout than the direct topological definition of intersection homology groups. It is tempting to try to define suitable generalizations of intersection homology groups in order to "explain" various results below (3.2, 3.4, 4.1, 5.2).

*Department of Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK. E-mail: b.totaro@dpmms.cam.ac.uk

2. The weight filtration

Deligne discovered a remarkable structure on the rational cohomology of any complex algebraic variety, not necessarily smooth or compact: the weight filtration [9]. This filtration expresses the way in which the cohomology of any variety is related to the cohomology of smooth compact varieties. It is a deep fact that the resulting filtration is well-defined. For example, an immediate consequence of the well-definedness of the weight filtration on cohomology with compact support is the following fact, originally conjectured by Serre ([11], [6], [12], p. 92).

Theorem 2.1. *For any complex algebraic variety X, not necessarily smooth or compact, one can define "virtual Betti numbers" $a_iX \in \mathbf{Z}$ for $i \geq 0$ such that*

(1) if X is smooth and compact, then the numbers a_iX are the Betti numbers $b_iX = \dim_{\mathbf{Q}} H^i(X, \mathbf{Q})$;

(2) for any Zariski-closed subset $Y \subset X$, $a_iX = a_iY + a_i(X - Y)$.

Using resolution of singularities, it is clear that the numbers a_iX are uniquely characterized by these properties. What is less clear is the existence of such numbers. It follows, for example, that if two smooth compact varieties X and Y can be written as finite disjoint unions of locally closed subsets, $X = \coprod X_i$ and $Y = \coprod Y_i$, with isomorphisms $X_i \cong Y_i$ for all i, then X and Y have the same Betti numbers. This is a topological property of algebraic varieties which has no obvious analogue in a purely topological context.

The existence of the weight filtration, and consequently of the virtual Betti numbers a_iX, was originally suggested by Grothendieck's approach to the Weil conjectures on counting rational points on varieties over finite fields. Indeed, the number of $\mathbf{F}_q$-points of a variety clearly has an additive property analogous to property (2) above. One proof of the existence of the weight filtration for complex varieties reduces the problem to the full Weil conjecture for varieties over finite fields, proved by Deligne [8]. Around the same time, Deligne gave a more direct proof of the existence of the weight filtration for complex varieties, using Hodge theory [7]. This is a classic example of the philosophy that the deepest properties of algebraic varieties can often be proved using either number theory or analysis, while they have no "purely geometric" proof.

In 1995, however, Gillet and Soulé gave a new proof of the existence of the weight filtration [13]. They used "only" resolution of singularities and algebraic K-theory, specifically the Gersten resolution. As a result of their more geometric proof, they were able to define the weight filtration on the integral cohomology or $\mathbf{F}_l$-cohomology of a complex algebraic variety, not only on rational cohomology.

To understand what this means, let me describe the weight filtration for a smooth complex variety U, not necessarily compact. Using resolution of singularities, we can write U as the complement of a divisor with normal crossings D in some smooth compact variety X. For $i \geq 0$, let $X^{(i)}$ be the disjoint union of the i-fold intersections of divisors. Then there is a spectral sequence

$$E_1 = H^j(X^{(i)}, k) \Rightarrow H_c^{i+j}(U, k)$$

for any coefficient ring k. The weight filtration on the compactly supported cohomology of U is defined as the filtration associated to this spectral sequence. Gillet

and Soulé show that for any coefficient ring k, this filtration is an invariant of U, independent of the choice of compactification U. This is not at all clear from the known invariance of this filtration for $k = \mathbf{Q}$.

In fact, Gillet and Soulé proved more: for any coefficient ring k, the spectral sequence is an invariant of U from the E_2 term on. For $k = \mathbf{Q}$, the spectral sequence degenerates at E_2, but this is not true with coefficients in $\mathbf{Z}$ or $\mathbf{F}_l$. As a result, for general coefficients k, the groups in the E_2 term are interesting new invariants of U which are not simply the associated graded groups to the weight filtration. They satisfy Mayer-Vietoris sequences, and so can be considered as a cohomology theory on algebraic varieties.

I can now explain a new application of the geometric proof that the weight filtration is well-defined. Namely, one can try to define the weight filtration not only for algebraic varieties. The point is that resolution of singularities holds more generally, for complex analytic spaces, and even for real analytic spaces. Gillet and Soulé's construction of the weight filtration uses algebraic K-theory as well as resolution of singularities, and it is not clear how to adapt the argument to an analytic setting. But Guillen and Navarro Aznar improved Gillet and Soulé's argument so as to construct the weight filtration using only resolution of singularities [17]. The details of their argument use their idea of "cubical hyperresolutions" [18].

Using the method of Guillen and Navarro Aznar, I have been able to define the weight filtration for complex and real analytic spaces. In more detail, let us define a compactification of a complex analytic space X to be a compact complex analytic space $\overline{X}$ containing X as the complement of a closed analytic subset. Of course, not every complex analytic space has a compactification in this sense. We say that two compactifications of X are equivalent if there is a third which lies over both of them.

Theorem 2.2. *Let k be any commutative ring. Then the compactly supported cohomology $H_c^*(X, k)$ has a well-defined weight filtration for every complex analytic space X with an equivalence class of compactifications.*

Any algebraic variety comes with a natural equivalence class of compactifications, but in the analytic setting this has to be considered as an extra piece of structure. On the other hand, the theorem says that the weight filtration is well-defined on all compact complex analytic spaces, with no extra structure needed.

For real analytic spaces, one has the difficulty that there is no natural orientation, unlike the complex analytic situation. This is not a problem if one uses $\mathbf{F}_2$-coefficients, and therefore one can prove:

Theorem 2.3. *For every real analytic space X with an equivalence class of compactifications, the compactly supported cohomology of the space $X(\mathbf{R})$ of real points with $\mathbf{F}_2$ coefficients has a well-defined weight filtration.*

In particular, one can define virtual Betti numbers $a_i X$ for a real analytic space X with an equivalence class of compactifications, the integers $a_i X$ being the usual $\mathbf{F}_2$-Betti numbers in the case of a closed real analytic manifold.

Example. Let X be the compact real analytic space obtained by identifying two copies at the circle at a point, and let Y be the compact real analytic space obtained by identifying two points on a single circle (the figure eight). It is imme-

diate to compute that $a_0X = 1$ and $a_1X = 2$, whereas $a_0Y = 0$ and $a_1 = Y$. The interesting point here is that the spaces $X(\mathbf{R})$ and $Y(\mathbf{R})$ of real points are homeomorphic. Thus the numbers a_i for a compact real analytic space are not topological invariants of the space of real points. In a similar vein, Steenbrink showed that the weight filtration on the rational cohomology of complex algebraic varieties is not a topological invariant, using 3-folds [27].

Nonetheless, it seems fair to say that extending the weight filtration and the virtual Betti numbers to complex and real analytic spaces helps to bring out more of the topological meaning of these invariants of algebraic varieties. A real analytic space has in some ways a weak structure; for example, the classification of closed real analytic manifolds up to isomorphism is the same as the classification of closed differentiable manifolds up to diffeomorphism. From this point of view, it is surprising that compactified real analytic spaces have the extra structure of the weight filtration on their $\mathbf{F}_2$-cohomology. It seems natural to ask for an $\mathbf{F}_2$-linear abelian category of "mixed motives" associated to compactified real analytic spaces X, such that the $\mathbf{F}_2$-cohomology groups of X with their weight filtration are determined by the mixed motive of X. On Beilinson's conjectured abelian category of mixed motives in algebraic geometry, see for example Jannsen [21], 11.3, and [22]; on various approximations to this category, see the triangulated categories defined by Hanamura [19], Levine [26], and Voevodsky [29], and the abelian category defined by Nori.

It should be much easier to define mixed motives for real analytic spaces than to do so for algebraic varieties. In particular, one might speculate that the mixed motive of a real analytic space should not involve much more information than the weight spectral sequence converging to its $\mathbf{F}_2$-cohomology (starting at E_2), perhaps considered together with an action of the Steenrod algebra. In low dimensions, one could hope for precise classifications of mixed motives along these lines.

3. Stringy Betti numbers

The following result of Batyrev's [4] is related to his famous result that two birational Calabi-Yau manifolds have the same Betti numbers. The proof uses Kontsevich's idea of motivic integration [24], as developed by Denef and Loeser [10]. To be precise, Batyrev's statement involves Hodge numbers, but I will only state what it gives about Betti numbers.

Theorem 3.1. *Let Y be a complex projective variety with log-terminal singularities. Then one can define the "stringy Poincaré function" $p_{str}(Y)$, which is a rational function, such that for any crepant resolution of singularities $\pi : X \to Y$, the stringy Poincaré function of Y is the usual Poincaré polynomial of X.*

We recall Reid's important definitions which are used here. First, let Y be any normal complex variety such that the canonical divisor K_Y is $\mathbf{Q}$-Cartier. By Hironaka, Y has a resolution of singularities $\pi : X \to Y$ such that the exceptional divisors E_i, $i \in I$, are smooth with normal crossings. The discrepancies a_i of E_i are defined by

$$K_X = \pi^* K_Y + \sum a_i E_i.$$

The variety Y is defined to have log-terminal singularities if and only if $a_i > -1$ for all i. A resolution $X \to Y$ is said to be crepant if $K_X = \pi^* K_Y$.

Batyrev defines the stringy Poincaré function of Y by the formula:

$$p_{\mathrm{str}}(Y) = \sum_{J \subset I} p(E_J^0) \prod_{j \in J} \frac{q-1}{q^{a_j+1}-1}.$$

Here E_J^0 is the open stratum of $E_J := \cap_{j \in J} E_j$, and $p(E_J^0)$ denotes the virtual Poincaré polynomial of E_J^0, written as a polynomial in $q^{1/2}$. Thus $p_{\mathrm{str}}(Y)$ is a rational function in $q^{1/2}$ for Y Gorenstein, and in $q^{1/n}$ for some n in general.

Batyrev's proof that the stringy Poincaré function of Y is independent of the choice of resolution, using motivic integration, rests on the additivity properties of the virtual Poincaré polynomial. Using our extension of virtual Betti numbers to complex analytic spaces, we find:

Theorem 3.2. *The stringy Poincaré function can be defined as a rational function for any compactified complex analytic space with log-terminal singularities. For any crepant resolution $X \to Y$ with Y compact, the stringy Poincaré function of Y is the usual Poincaré polynomial of X.*

Likewise for real analytic spaces:

Theorem 3.3. *An $\mathbf{F}_2$-analogue of the stringy Poincaré function can be defined as a rational function for compactified real analytic spaces with log-terminal singularities. For any crepant resolution $X \to Y$ with Y compact, the stringy Poincaré function of Y is the usual Poincaré polynomial of the $\mathbf{F}_2$-cohomology of X.*

In particular, this answers part of Goresky and MacPherson's Problem 7 in [15]:

Corollary 3.4. *Given a compact real algebraic variety Y, the $\mathbf{F}_2$-Betti numbers of any two projective IH-small resolutions of Y are the same.*

This uses the relation between IH-small resolutions and crepant resolutions, which I worked out in [28] using results of Kawamata [23] and Wisniewski [30]. In the complex situation, the corollary (for Betti numbers with any coefficients) has a more direct proof, since the Betti numbers of any small resolution of Y are equal to the dimensions of the intersection homology groups of Y. It is not yet known whether one can define a new version of intersection homology groups with $\mathbf{F}_2$-coefficients which would be self-dual for all compact real analytic spaces. A possible framework for defining such a theory has been set up by Banagl [2].

4. The elliptic genus of a singular variety

I found that any characteristic number which can be extended from smooth compact complex varieties to singular varieties, compatibly with small resolutions, must be a specialization of the elliptic genus [28]. It was then an important problem to define the elliptic genus for singular varieties. This was solved in a completely satisfying way by Borisov and Libgober [5]:

Theorem 4.1. *Let Y be a projective variety with log-terminal singularities. Then one can define the elliptic genus of Y, $\varphi(Y)$, such that for any crepant resolution $X \to Y$, we have $\varphi(Y) = \varphi(X)$.*

Here is Borisov and Libgober's definition of $\varphi(Y)$. Let $\pi : X \to Y$ be a resolution whose exceptional divisors E_k have simple normal crossings, and let a_k be the discrepancies as in section 3. Formally, let y_l denote the Chern roots of X so that $c(TX) = \prod_l (1 + y_l)$, and let e_k be the cohomology classes on X of the divisors E_k. Then $\varphi(Y)$ is the analytic function of variables z and τ defined by

$$\varphi(Y) = \int_Y (\prod_l \frac{(\frac{y_l}{2\pi i})\theta(\frac{y_l}{2\pi i} - z)\theta'(0)}{\theta(-z)\theta(\frac{y_l}{2\pi i})}) \times (\prod_k \frac{\theta(\frac{e_k}{2\pi i} - (\alpha_k + 1)z)\theta(-z)}{\theta(\frac{e_k}{2\pi i} - z)\theta(-(\alpha_k + 1)z)}),$$

where $\theta(z, \tau)$ is the Jacobi theta function. The proof that $\varphi(Y)$ is independent of the choice of resolution for log-terminal Y uses the weak factorization theorem of Abramovich, Karu, Matsuki, and Wlodarczyk ([1], [31]).

In the spirit of earlier sections, the singular elliptic genus extends to compact complex analytic spaces with log-terminal singularities. But it remains a mystery how to define the elliptic genus for some topologically defined class of singular spaces that would include singular analytic spaces with log-terminal singularities.

5. Possible characteristic numbers for real analytic spaces

In my paper [28], in trying to define characteristic numbers for singular complex varieties, it was very helpful to require that these numbers are compatible with IH-small resolutions, as Goresky and MacPherson had suggested ([15], Problem 10). The problem thereby becomes more precise: it may be possible to show that some characteristic numbers extend to singular varieties and some do not. This can help to suggest valuable invariants for singular varieties, such as Borisov and Libgober's elliptic genus for singular varieties, even if one is not a priori interested in IH-small resolutions. (The same comments apply to crepant resolutions.)

With this in mind, we here begin to analyze which characteristic numbers can be defined for real analytic spaces, or for topological spaces with similar singularities, compatibly with IH-small resolutions. In the complex situation, the fundamental example of a singularity with two different IH-small resolutions is the 3-fold node; one says that the two IH-small resolutions are related by the simplest type of "flop." Likewise, in the real situation, the real 3-fold node has two different IH-small resolutions. For convenience, let us say that two closed n-manifolds are related by a "real flop" if they are the two different IH-small resolutions X_1 and X_2 of a singular space with singular set of real codimension 3 that is locally isomorphic to the product of the 3-fold node with an $(n-3)$-manifold.

Let us first consider characteristic numbers for unoriented spaces. By Thom, the bordism ring MO_* for unoriented manifolds is detected by Stiefel-Whitney numbers. Therefore we can ask which Stiefel-Whitney numbers (meaning $\mathbf{F}_2$-linear combinations of Stiefel-Whitney monomials) are unchanged under real flops. Or, more or less equivalently: what is the quotient of the bordism ring MO_* by the ideal of real flops $X_1 - X_2$, for X_1 and X_2 as above? There is a good answer:

Theorem 5.1. *The $\mathbf{F}_2$-vector space of Stiefel-Whitney numbers which are invariant under real flops of n-manifolds is spanned by the numbers $w_1^i w_{n-i}$ for $0 \leq i \leq n$, or equivalently by the numbers $w_1^{n-2i} v_i^2$ for $0 \leq i \leq n/2$, modulo those Stiefel-Whitney numbers which vanish for all n-manifolds. Here $v_i = v_i(w_1, w_2, \ldots)$ denotes the Wu class. The dimension of this space of invariant Stiefel-Whitney numbers, modulo those which vanish for all n-manifolds, is 0 for n odd and $\lfloor n/2 \rfloor + 1$ for n even. The quotient ring of MO_* by the ideal of real flops is isomorphic to:*

$$\mathbf{F}_2[\mathbf{RP}^2, \mathbf{RP}^4, \mathbf{RP}^8, \ldots]/((\mathbf{RP}^{2^a})^2 = (\mathbf{RP}^2)^{2^a} \text{ for all } a \geq 2).$$

This class of Stiefel-Whitney numbers has occurred before, in Goresky and Pardon's calculation of the bordism ring of locally orientable $\mathbf{F}_2$-Witt spaces [16]. To be precise, the latter ring coincides with the above ring in even dimensions but is also nonzero in odd dimensions. Goresky defined a Wu class v_i in intersection homology for $\mathbf{F}_2$-Witt spaces [14], so that the square v_i^2 lives in ordinary homology, and the characteristic numbers for locally orientable $\mathbf{F}_2$-Witt spaces Y are obtained by multiplying these homology classes by powers of the cohomology class w_1.

This does not explain the invariance of these Stiefel-Whitney numbers for real flops, however. The problem is that the 3-fold node is not an $\mathbf{F}_2$-Witt space. (Topologically, it is the cone over $S^1 \times S^1$, whereas the cone over an even-dimensional manifold is a Witt space if and only if the homology in the middle dimension is zero.) That is, the standard definition of intersection homology is not self-dual on a space with 3-fold node singularities. This again points to the problem of defining a new version of intersection homology with $\mathbf{F}_2$ coefficients which is self-dual on real analytic spaces. That should yield an L-class in the $\mathbf{F}_2$-homology of such a space, which we can also identify with the square of the Wu class, and which therefore should allow the above characteristic numbers to be defined for a large class of real analytic spaces. There are related results by Banagl [3], for spaces which admit an extra "Lagrangian" structure.

We now ask the analogous question for oriented singular spaces: what characteristic numbers can be defined, compatibly with IH-small resolutions? We could begin by asking for the quotient ring of the oriented bordism ring MSO_* by oriented real flops $X_1 - X_2$, defined exactly as in the unoriented case (X_1 and X_2 are the two small resolutions of a family of real 3-fold nodes), except that we require X_1 and X_2 to be compatibly oriented. It turns out that this is not enough: all Pontrjagin numbers are invariant under oriented real flops, whereas they can change under other changes from one IH-small resolution to another, such as complex flops (between the two small resolutions of a complex family of complex 3-fold nodes). By considering both real and complex flops, we get a reasonable answer:

Theorem 5.2. *The quotient ring of MSO_* by the ideal generated by oriented real flops and complex flops is:*

$$\mathbf{Z}[\delta, 2\gamma, 2\gamma^2, 2\gamma^4, \ldots],$$

where $\mathbf{CP}^2$ maps to δ and $\mathbf{CP}^4$ maps to $2\gamma + \delta^2$. This quotient ring is exactly the image of MSO_* under the Ochanine elliptic genus ([25], p. 63).

This result suggests that it should be possible to define the Ochanine genus for a large class of compact oriented real analytic spaces, or even more general singular spaces.

References

[1] D. Abramovich, K. Karu, K. Matsuki, and J. Wlodarczyk, Torification and factorization of birational maps, math.AG/9904135.

[2] M. Banagl, Extending intersection homology type invariants to non-Witt spaces, *Memoirs of the AMS*, to appear.

[3] M. Banagl, The L-class of non-Witt spaces, to appear.

[4] V. Batyrev, Stringy Hodge numbers of varieties with Gorenstein canonical singularities, *Integrable systems and algebraic geometry (Kobe/Kyoto, 1997)*, 1–32, World Scientific, 1998.

[5] L. Borisov and A. Libgober, Elliptic genera of singular varieties, *Duke Math. J.*, to appear.

[6] V. Danilov and A. Khovanskii, Newton polyhedra and an algorithm for computing Hodge-Deligne numbers, *Math. USSR Izv.* **29** (1987), 279–298.

[7] P. Deligne, Théorie de Hodge I, II, III, *Proc. ICM 1970*, v. 1, 425–430; *Publ. Math. IHES* **40** (1972), 5–57; **44** (1974), 5–78.

[8] P. Deligne, La conjecture de Weil I, *Publ. Math. IHES* **43** (1974), 273–308.

[9] P. Deligne, Poids dans la cohomologie des variétés algébriques, Actes ICM Vancouver 1974, I, 79–85.

[10] J. Denef and F. Loeser, Germs of arcs on singular algebraic varieties and motivic integration, *Invent. Math.* **135** (1999), 201–232.

[11] A. Durfee, Algebraic varieties which are a disjoint union of subvarieties, *Geometry and topology: manifolds, varieties and knots*, 99–102, Marcel Dekker, 1987.

[12] W. Fulton, Introduction to toric varieties, Princeton, 1993.

[13] H. Gillet and C. Soulé, Descent, motives, and K-theory, *J. reine angew. Math.* **478** (1996), 127–176.

[14] M. Goresky, Intersection homology operations, *Comment. Math. Helv.* **59** (1984), 485–505.

[15] M. Goresky and R. MacPherson, Problems and bibliography on intersection homology, *Intersection homology*, ed. A. Borel, Birkhäuser, 1984, 221–233.

[16] M. Goresky and W. Pardon, Wu numbers of singular spaces, *Topology* **28** (1989), 325–367.

[17] F. Guillén and V. Navarro Aznar, Un critère d'extension d'un foncteur défini sur les schémas lisses, math.AG/9505008.

[18] F. Guillén, V. Navarro Aznar, P. Pascual, and F. Puerta, *Hyperrésolutions cubiques et descente cohomologique*, Lecture Notes in Mathematics 1335, Springer, 1988.

[19] M. Hanamura, Homological and cohomological motives of algebraic varieties, *Invent. Math.* **142** (2000), 319–349.

[20] H. Hironaka, Resolution of singularities of an algebraic variety over a field of

characteristic zero, *Ann. Math.* **79** (1964), 109–326.
[21] U. Jannsen, *Mixed motives and algebraic K-theory,* LNM 1400, Springer, 1990.
[22] U. Jannsen, Motivic sheaves and filtrations on Chow groups, *Motives (Seattle, 1991)*, 245–302, AMS, 1994.
[23] Y. Kawamata, K. Matsuda, and K. Matsuki, Introduction to the minimal model program, *Algebraic Geometry (Sendai, 1985)*, ed. T. Oda, 283–360, Kinokuniya-North Holland, 1987.
[24] M. Kontsevich, lecture at Orsay, 7 December 1995.
[25] P. Landweber, Elliptic cohomology and modular forms, *Elliptic curves and modular forms in algebraic topology,* 55–68, LNM 1326, Springer, 1988.
[26] M. Levine, *Mixed motives,* AMS, 1998.
[27] J. Steenbrink, Topological invariance of the weight filtration, *Indag. Math.* **46** (1984), 63–76.
[28] B. Totaro, Chern numbers for singular varieties and elliptic homology, *Ann. Math.* **151** (2000), 757–791.
[29] V. Voevodsky, A. Suslin, and E. Friedlander, *Cycles, transfers, and motivic homology theories,* Princeton, 2000.
[30] J. Wisniewski, On contractions of extremal rays of Fano manifolds, *J. reine angew. Math.* **417** (1991), 141–157.
[31] J. Wlodarczyk, Combinatorial structures on toroidal varieties and a proof of the weak factorization theorem, math.AG/9904076.

Section 7. Lie Group and Representation Theory

Patrick Delorme: *Harmonic Analysis on Real Reductive Symmetric Spaces* 545

Pavel Etingof: *On the Dynamical Yang-Baxter Equation* 555

D. Gaitsgory: *Geometric Langlands Correspondence for GL_n* 571

Michael Harris: *On the Local Langlands Correspondence* 583

Alexander Klyachko: *Vector Bundles, Linear Representations, and Spectral Problems* 599

Toshiyuki Kobayashi: *Branching Problems of Unitary Representations* 615

Vikram Bhagvandas Mehta: *Representations of Algebraic Groups and Principal Bundles on Algebraic Varieties* 629

E. Meinrenken: *Clifford Algebras and the Duflo Isomorphism* 637

Maxim Nazarov: *Representations of Yangians Associated with Skew Young Diagrams* 643

Freydoon Shahidi: *Automorphic L-Functions and Functoriality* 655

Marie-France Vignéras: *Modular Representations of p-adic Groups and of Affine Hecke Algebras* 667

ICM 2002 · Vol. II · 545–554

Harmonic Analysis on Real Reductive Symmetric Spaces

Patrick Delorme*

Abstract

Let G be a reductive group in the Harish-Chandra class e.g. a connected semisimple Lie group with finite center, or the group of real points of a connected reductive algebraic group defined over $\mathbb{R}$. Let σ be an involution of the Lie group G, H an open subgroup of the subgroup of fixed points of σ. One decomposes the elements of $L^2(G/H)$ with the help of joint eigenfunctions under the algebra of left invariant differential operators under G on G/H.

2000 Mathematics subject classification: 22E46, 22F30, 22E30, 22E50, 33C67.
Keywords and Phrases: Reductive symmetric space, Plancherel formula, Meromorphic continuation of Eisenstein integrals, Temperedness, Truncation, Maass-Selberg relations.

1. Introduction

Let G be a real reductive group in the Harish-Chandra class [H-C1], e.g. a connected semisimple Lie group with finite center, or the group of real points of a connected reductive algebraic group defined over $\mathbb{R}$. Let σ be an involution of the Lie group G, H an open subgroup of the subgroup of fixed points of σ.

Important problems of harmonic analysis on the so-called reductive symmetric space G/H are :

(a) to make the simultaneous spectral decomposition of the elements of the algebra $\mathbb{D}(G/H)$ of left invariant differential operators under G on G/H. In other words, one wants to write the elements of $L^2(G/H)$ with the help of joint eigenfunctions under $\mathbb{D}(G/H)$.

(b) to decompose the left regular representation of G in $L^2(G/H)$ into an Hilbert integral of irreducible unitary representations of G : this is essentially the Plancherel formula.

*Institut de Mathématiques de Luminy, U.P.R. 9016 du C.N.R.S., Faculté des Sciences de Luminy, 163 Avenue de Luminy, Case 930, 13288 Marseille Cedex 09, France. E-mail: delorme@iml.univ-mrs.fr

(c) to decompose the Dirac measure at eH, where e is the neutral element of G, into an integral of H-fixed distribution vectors : this is essentially the Fourier inversion formula.

These problems were solved for the "group case" (i.e. the group viewed as a symmetric space : $G = G_1 \times G_1$, $\sigma(x,y) = (y,x)$, H is the diagonal of $G_1 \times G_1$) by Harish-Chandra in 1970s (see [H-C1,2, 3]), the Riemannian case (H maximal compact) had been treated before (see [He]). Later, there were deep results by T. Oshima [O1]. When G is a complex group and H is a real form, the Problems (a), (b), (c) were solved by P. Harinck, together with an inversion formula for orbital integrals ([Ha], see also [D3] for the link of her work with the work of A. Bouaziz on real reductive groups).

Then, E. van den Ban and H. Schlichtkrull, on the one hand, and I, on the other hand, obtained different solutions to problems (a), (b), (c). Moreover, they obtained a Paley-Wiener theorem (see [BS3] for a presentation of their work). I present here my point of view, with an emphasize on problem (a), because it simplifies the formulations of the results (nevertheless, the important aspect of representation theory is hidden). It includes several joint works, mainly with J. Carmona , and also with E. van den Ban and J.L. Brylinski. Severals works of T. Oshima, linked to the the Flensted-Jensen duality, alone and with T. Matsuki are very important in my proof, as well as earlier results of E. van den Ban and H. Schlichtkrull.

I have to acknowledge the deep influence of Harish-Chandra 's work. The crucial role played by the work [Be] of J. Bernstein on the support of the Plancherel measure, and some part of Arthur's article on the local trace formula [A] will be appearant in the main body of the article.

2. Temperedness of the spectrum

Let θ be a Cartan involution of G commuting with σ, let K be the fixed point set of θ. Let $\mathfrak{g}$ be the Lie algebra of G, etc. Let $\mathfrak{s}$ (resp. $\mathfrak{q}$) be the space of elements in $\mathfrak{g}$ which are antiinvariant under the differential of θ (resp. σ). Let $\mathfrak{a}_\emptyset$ be a maximal abelian subspace of $\mathfrak{s} \cap \mathfrak{q}$. If P is a $\sigma\theta$-stable parabolic subgroup of G, containing $A_\emptyset := exp\, \mathfrak{a}_\emptyset$, we denote by $P = M_P A_P N_P$ its Langlands σ-decomposition. More precisely A_P is the subgroup of the elements a of the split component of the Levi factor $L_P = P \cap \theta(P)$ such that $\sigma(a) = a^{-1}$. Here M_P is larger than that for the usual Langlands decomposition.

In order to simplify the exposition we will make the following simplifying assumption:

Hypothesis: *For any P as above, HP is the unique open (H,P)-double coset.*

When $\sigma = \theta$ (the case of a riemannian symmetric space) or the "group case", this hypothesis is satisfied.

To get the Plancherel formula, it is useful to use K-finite functions. They are often replaced by τ-spherical functions. Here (τ, V_τ) is a finite dimensional unitary representation of K and a τ-spherical function on G/H is a function $f : G/H \to V_\tau$ such that $f(kx) = \tau(k)f(x)$, $k \in K$, $x \in G/H$.

Some spaces of τ-spherical functions on G/H play a crucial role in the theory, namely:

(a) $\mathcal{C}(G/H, \tau)$: the Schwartz space of τ-spherical functions on G/H which are rapidly decreasing as well as their derivatives by elements of the enveloping algebra $U(\mathfrak{g})$ of $\mathfrak{g}$ (see [B2]).

(b) $\mathcal{A}(G/H, \tau)$: the space of smooth τ-spherical functions on G/H which are $\mathbb{D}(G/H)$ finite. Here $\mathcal{A}$ is used to evoke automorphic forms.

(c) $\mathcal{A}_{temp}(G/H, \tau)$: the space of elements of $\mathcal{A}(G/H, \tau)$ which have tempered growth as well as their derivatives by elements of $U(\mathfrak{g})$ ([D2]). Integration of functions on G/H defines a pairing between $\mathcal{A}_{temp}(G/H, \tau)$ and $\mathcal{C}(G/H, \tau)$.

(d) $\mathcal{A}_2(G/H, \tau)$: the space of square integrable elements of $\mathcal{A}(G/H, \tau)$. This is a subspace of the three preceeding spaces.

One has:

Theorem 1 ([D2]): *The space $\mathcal{A}_2(G/H, \tau)$ is finite dimensional.*

This is deduced from the theory of discrete series for G/H initiated by M. Flensted-Jensen [F-J] and achieved by T. Oshima and T. Matsuki, using the Flensted-Jensen duality [OM]. One has also to use the behaviour of the discrete series under certain translation functors, studied by D. Vogan [V] and a result of H. Schlichtkrull [S] on the minimal K-types of certain discrete series.

The next result follows from the work of J. Bernstein [Be] on the support of the Plancherel measure.

Theorem 2 ([CD1], Appendice C): *Every function in $\mathcal{C}(G/H, \tau)$ can be canonically desintegrated as an integral of elements of $\mathcal{A}_{temp}(G/H, \tau)$.*

This information appeared to be crucial at the end of our proof.

3. The continuous spectrum: Eisenstein integrals

Let $P = MAN$ the Langlands σ-decomposition of a $\sigma\theta$-stable parabolic subgroup P of G. Let ρ_P be the half sum of the roots of $\mathfrak{a}$ in $\mathfrak{n}$ and $\lambda \in \mathfrak{a}_{\mathbb{C}}^*$ be such that the real part of $\lambda - \rho_P$ is strictly dominant with respect to the roots of $\mathfrak{a}$ in $\mathfrak{n}$. Let τ_M be the restriction of τ to $M \cap K$. Then, if $x \in G/H$ and $\psi \in \mathcal{A}_2(M/M \cap H, \tau_M)$, the following integral is convergent:

$$E(P, \psi, \lambda)(x) := \int_K \tau(k^{-1}) \Psi_\lambda(kx) dk,$$

where $\Psi_\lambda(x) = 0$ if $x \notin PH$, and $\Psi_\lambda(x) = a^{-\lambda+\rho_P}\psi(m)$ if $x = namH$ with $n \in N$, $a \in A$, $m \in M$. Moreover $E(P, \psi, \lambda)$ is an element of $\mathcal{A}(G/H, \tau)$. Eisenstein integrals are the τ-spherical versions of K-finite functions of the form: $gH \mapsto < \pi'(g)\xi, v >$, where π' is the contragredient of a generalized principal series π, ξ is a certain H-fixed distribution vector of π, v is a K-finite vector of π.

Theorem 3 ([BrD]): *The function $\lambda \mapsto E(P, \psi, \lambda)$ admits a meromorphic continuation in $\lambda \in \mathfrak{a}_{\mathbb{C}}^*$. This meromorphic continuation, denoted in the same way, multiplied by a suitable product, p_ψ, of functions of type $\lambda \mapsto (\alpha, \lambda) + c$, where α is a root of $\mathfrak{a}$ and $c \in \mathbb{C}$, is holomorphic around $i\mathfrak{a}^*$.*

This meromorphic continuation is an interesting feature of the theory. For the the "group case", it comes down to the meromorphic continuation of Knapp-Stein intertwining integrals. My proof with Brylinski uses D-modules arguments.

The case where P is minimal had been treated separately by E. van den Ban [B1] and G. Olafsson [Ol]. One has also to mention the earlier work of T. Oshima and J. Sekiguchi [OSe] on the spaces of type G/K_ε.

The proof which gives the best results uses a method of tensoring by finite dimensional representations. It is a joint work with J. Carmona. It was initiated by D. Vogan and N. Wallach (see [W], chapter 10) for the meromorphic continuation of the Knapp-Stein intertwining integrals. For symmetric spaces and the most continuous spectrum, the proof is due to E. van den Ban [B2]. This proof uses Bruhat's thesis and tensoring by finite dimensional modules. This implies rough estimates for Eisenstein integrals, which generalize those obtained by E. van den Ban when P is minimal [B2].

Theorem 4 ([D1]): *If $\lambda \in i\mathfrak{a}^*$ is such that $E(P,\psi,\lambda)$ is well defined, then it is tempered, i.e. is an element of $\mathcal{A}_{temp}(G/H,\tau)$.*

This is a natural result but the proof is quite long. It uses the behaviour under translation functors of H-fixed distribution vectors of discrete series and of generalized principal series, and also of the Poisson transform. Moreover the duality of M. Flensted-Jensen, [F-J], and a criteria of temperedness due to Oshima [O2] play a crucial role (apparently, J. Carmona has a way to avoid boundary values).

With the help of this theorem and by using techniques due to E. van den Ban [B2], the rough estimates for Eisenstein integrals can be improved to get uniform sharp estimates for $p_\psi(\lambda)E(P,\psi,\lambda)$, $\lambda \in i\mathfrak{a}^*$ (cf. [D1]).

4. C-functions

Let P be as above and let L be equal to MA. The theory of the constant term, due to J. Carmona [C1] (Harish-Chandra for the group case, [H-C1]), gives a linear map from $\mathcal{A}_{temp}(G/H,\ \tau)$ into $\mathcal{A}_{temp}(L/L\cap H,\ \tau_L)$, $\varphi \mapsto \varphi_P$, characterized by :

$$lim_{t\to+\infty}\ \delta_P^{1/2}((exp\,tX)l)\varphi((exp\,tX)l) - \varphi_P((exp\,tX)l) = 0,$$

where $l \in L/L\cap H$, $X \in \mathfrak{a}_P$ is P-dominant and δ_P is the modular function of P.

Let Q be a $\sigma\theta$-stable parabolic subgroup of G with the same θ-stable Levi subgroup L other than P. Let $W(\mathfrak{a})$ be the group of automorphisms of $\mathfrak{a}$ induced by an element of $Ad(G)$. One defines meromorphic functions on $\mathfrak{a}_{\mathbb{C}}^*$, $\lambda \mapsto C_{Q|P}(s,\lambda)$ with values in $End(\mathcal{A}_2(M/M\cap H,\tau_M))$ such that :

$$E(P,\psi,\lambda)_Q(ma) = \sum_{s\in W(\mathfrak{a})} (C_{Q|P}(s,\lambda)\psi)(m)a^{-s\lambda},\ m\in M,\ a\in A,\ \lambda\in i\mathfrak{a}^*,$$

or rather for λ in an open dense subset of $i\mathfrak{a}^*$.

The C-functions allow to normalize Eisenstein integrals as folllows:

$$E^0(P,\psi,\lambda) := E(P, C_{P|P}(1,\lambda)^{-1}\psi,\lambda).$$

5. Truncation, Maass-Selberg relations and the regularity of normalized Eisenstein integrals

Let P be as above and let $P' = M'A'N'$ be the Langlands σ-decomposition of another $\sigma\theta$-stable parabolic subgroup of G. Let ψ (resp. ψ') be an element of $\mathcal{A}_2(M/M\cap H, \tau_M)$ (resp. $\mathcal{A}_2(M'/M'\cap H, \tau_{M'})$). One chooses p_ψ as in Theorem 3, such that the product of p_ψ with the C-functions are holomorphic in a neighbourhood of $i\mathfrak{a}^*$ which is a product of $i\mathfrak{a}^*$ with a neighbourhood of 0 in $\mathfrak{a}^*$. We do the same for ψ'. One defines $F(\lambda) := p_\psi(\lambda)E(P,\psi,\lambda)$. One defines similarly F'.

One assumes, for the rest of the article, that G is semisimple. This is just to simplify the exposition. One chooses $T \in \mathfrak{a}_\emptyset$, regular with respect to the roots of $\mathfrak{a}_\emptyset$ in $\mathfrak{g}$. Let C_T^1 be the convex hull of $W(\mathfrak{a}_\emptyset)T$ and let C_T be equal to the subset $K(exp\, C_T^1)H$ of G/H.

Theorem 5 ([D2])**:**

(i) *One gets an explicit expression $\omega^T(\lambda,\lambda')$, involving the C-functions (see an example below) and vanishing when A and A' are not conjugate under K, which is asymptotic to*

$$\Omega^T(\lambda,\lambda') := \int_{C_T} (F(\lambda)(x), F'(\lambda')(x))dx,$$

when T goes to infinity and $\lambda \in i\mathfrak{a}^*$, $\lambda' \in i\mathfrak{a}'^*$. More precisely for $\delta > 0$ there exists $C > 0$, $k \in \mathbb{N}$ and $\varepsilon > 0$, such that for all T satisfying $\mid \alpha(T) \mid \geq \delta \mid\mid T \mid\mid$ for every root α of $\mathfrak{a}_\emptyset$ in $\mathfrak{g}$, one has:

$$\mid \Omega^T(\lambda,\lambda') - \omega^T(\lambda,\lambda') \mid \leq C(1+ \mid\mid \lambda \mid\mid)^k(1+ \mid\mid \lambda' \mid\mid)^k e^{-\varepsilon||T||}.$$

(ii) *Moreover ω^T is analytic in $(\lambda,\lambda') \in i\mathfrak{a}^* \times i\mathfrak{a}'^*$.*

This generalizes a result of J. Arthur for the group case [A], Theorem 8.1. My proof is quite similar, but I was able to avoid his use of the Plancherel formula.

(ii) is an easy consequence of (i). In fact, the explicit form of ω^T implies that it is meromorphic around $i\mathfrak{a}^* \times i\mathfrak{a}'^*$. Moreover Ω^T is holomorphic, hence locally bounded, around $i\mathfrak{a}^* \times i\mathfrak{a}'^*$. From the inequality in (i), one deduces that ω^T is locally bounded, hence holomorphic, around $i\mathfrak{a}^* \times i\mathfrak{a}'^*$.

We will now show, by an example, how the explicit form of ω^T and its analyticity in $(\lambda,\lambda') \in i\mathfrak{a}^* \times i\mathfrak{a}'^*$ imply the Maass-Selberg relations.

Let σ be equal to θ, $H = K$, and τ be the trivial representation. Let P, P' be minimal parabolic subgroups of G. Then $dim\mathcal{A}_2(M/M\cap H,\tau) = 1$ and the C-functions are scalar valued. One assumes $\mathfrak{g}$ to be semisimple and that the dimension of A is one. Then $W(\mathfrak{a})$ has 2 elements, ± 1, and one has the following explicit expression of ω^T:

$$\omega^T(\lambda,\lambda') = p_\psi(\lambda)p_{\psi'}(\lambda') \sum_{s=\pm 1, s'=\pm 1} e^{s\lambda T - s'\lambda' T} C_{P|P}(s,\lambda)\overline{C_{P|P}(s',\lambda')}(s\lambda - s'\lambda')^{-1}.$$

Thus $\omega^T(\lambda,\lambda')$ is the sum of a product of $(\lambda-\lambda')^{-1}$ by an analytic function with a product of $(\lambda+\lambda')^{-1}$ by an analytic function. The analyticity at (λ,λ) implies

easily that the factor in front of $(\lambda - \lambda')^{-1}$ vanishes for $\lambda = \lambda'$. Hence we get $\mid C_{P\mid P}(1,\lambda) \mid^2 = \mid C_{P\mid P}(-1,\lambda) \mid^2$, $\lambda \in i\mathfrak{a}^*$. This is one of the Maass-Selberg relations(cf. [D2], Theorem 2, and the work with J. Carmona [CD2], Theorem 2 for the general case, see [B1], [B2] for the case where P is minimal). These relations imply that the C-functions attached to normalized Eisenstein integrals are unitary, when defined, for λ purely imaginary. Hence they are locally bounded . This implies that they are holomorphic around the imaginary axis. This implies in particular some holomorphy property of the constant term of normalized Eisentein integrals. From this, with the help of [BCD], one deduces:

Theorem 6 (Regularity theorem for normalized Eisenstein integrals, [CD2], [BS1] for P minimal)**:** *The normalized Eisenstein integrals are holomorphic in a neighbourhood of the imaginary axis.*

6. Fourier transform and wave packets

Theorem 7 ([CD2], [BS1] for $\boldsymbol{P}$ minimal)**:** *For $f \in \mathcal{C}(G/H,\tau)$, one has $\mathcal{F}_P^0 f \in \mathcal{S}(i\mathfrak{a}^*) \otimes \mathcal{A}_2(M/M\cap H,\tau_M)$, where $\mathcal{F}_P^0 f$ is characterized by:*

$$((\mathcal{F}_P^0 f)(\lambda),\psi) = \int_{G/H} (f(x), E^0(P,\psi,\lambda)(x))dx, \quad \lambda \in i\mathfrak{a}^*, \quad \psi \in \mathcal{A}_2(M/M\cap H,\tau_M),$$

here $\mathcal{S}(i\mathfrak{a}^*)$ is the usual Schwartz space.

This theorem follows from the sharp estimates of Eisenstein integrals.

Theorem 8 ([BCD])**:** *If Ψ is an element of $\mathcal{S}(i\mathfrak{a}^*) \otimes \mathcal{A}_2(M/M\cap H,\tau_M)$, one has $\mathcal{J}_P^0 \in \mathcal{C}(G/H,\tau)$, where :*

$$\mathcal{J}_P^0(x) := \int_{i\mathfrak{a}^*} E^0(P,\Psi(\lambda),\lambda)d\lambda, \quad x \in G/H.$$

This theorem follows from the regularity theorem and from the joint work with E. van den Ban and J. Carmona, [BCD].

Now we want to compute $\mathcal{F}_{P'}^0 \mathcal{J}_P^0$. For this purpose one has to study the integral:

$$I := \int_{G/H} (\int_{i\mathfrak{a}^*} \alpha(\lambda) E^0(P,\psi,\lambda)(x)d\lambda, E^0(P',\psi',\lambda')(x))dx.$$

One truncates the integral on G/H to C_T and let T goes to infinity (far from the walls). Let us denote the truncated inner product of the normalized Eisenstein integrals by ${}^0\Omega^T(\lambda,\lambda')$. Using Fubini's theorem one has:

$$I = lim_{T\to\infty} \int_{i\mathfrak{a}^*} \alpha(\lambda) {}^0\Omega^T(\lambda,\lambda')d\lambda.$$

As for unnormalized Eisentein integrals, one has an asymptotic evaluation of ${}^0\Omega^T(\lambda,\lambda')$ by an explicit expression ${}^0\omega^T(\lambda,\lambda')$. One can replace ${}^0\Omega^T$ by ${}^0\omega^T$ in the previous formula. By using an expression of ${}^0\omega^T(\lambda,\lambda')$, viewed as a distribution in λ, for fixed λ' , involving Fourier transforms of cones ([D2], Theorem 3) one gets:

Theorem 9 ([CD2]): *Let $\mathbb{F}$ be a set of representative of σ-association classes of $\sigma\theta$-stable parabolic subgroups. Here σ-association means that the $\mathfrak{a}$ are conjugated under K. Define:*

$$\mathcal{P}_\tau = \sum_{P\in\mathbb{F}}((W(\mathfrak{a}_P))^{-1}\mathcal{J}_P^0\mathcal{F}_P^0.$$

Then $\mathcal{P}_\tau$ is an orthogonal projection operator in $\mathcal{C}(G/H,\tau)$ endowed with the L^2 scalar product.

7. The Plancherel formula

Essentially, the solution to problem (a) is contained in the following:

Theorem 10 ([D4]): *The projection $\mathcal{P}_\tau$ is the identity operator on $\mathcal{C}(G/H,\tau)$.*

Actually, this gives an expression of every element in $\mathcal{C}(G/H,\tau)$ as a wave packet of normalized Eisenstein integrals. The proof goes as follows. If $\mathcal{P}_\tau$ was not the identity, using Theorem 1 on the temperedness of the spectrum, one could find a non zero element of $\mathcal{A}_{temp}(G/H,\tau)$ which is orthogonal to the image of $\mathcal{P}_\tau$. Then, generalizing Theorem 5 to the truncated inner product of an Eisenstein integral with a general element of $\mathcal{A}_{temp}(G/H,\tau)$, this orthogonality can be explicitely described (cf. the evaluation of I before Theorem 9). As a result, this function has to be zero, a contradiction which proves the theorem.

The theorem translates to K-finite functions, involving representations and H-fixed distribution vectors.

The theorem can also be expressed with the unnormalized Eisenstein integrals. Then there are certain Plancherel factors involved. They are linked, as in the group case , to the intertwining integrals. Following the approach of A.Knapp and G.Zuckerman, [KZ], their computation is reduced to find an embedding of discrete series into principal series attached to minimal parabolic subgroups. For connected groups, this has been done by J. Carmona [C2].

8. Applications and open problems

Schwartz space for the hypergeometric Fourier transform

The image of a natural Schwartz space by the hypergeometric Fourier transform is characterized [D5]. The work uses the Plancherel formula of E. Opdam [Op], and the techniques mentioned above : theory of the constant term, C-functions, truncation ...

Generalized Schur orthogonality relations

Using the Plancherel formula for reductive symmetric spaces groups, K. Ankabout, [An], has proved generalized Schur orthogonality relations for generalized coefficients related to real reductive symmetric spaces. In particular, at least if we assume multiplicity one in the Plancherel formula, it implies the following:

There exists an explicit positive function d on G/H, such that, for almost all representations π occuring in the Plancherel formula, for ξ an H-fixed distribution vector of π occuring in the decomposition of the Dirac measure, there exists an

explicit non zero constant C_π such that, for all v, v', K-finite vectors in the space of π:

$$\lim_{\varepsilon \to 0^+} \varepsilon^{n_\pi} \int_{G/H} e^{-\varepsilon d(x)} < \pi'(g)\xi, v > \overline{< \pi'(g)\xi, v' >} dx = C_\pi(v, v').$$

Here n_π is the dimension of the support of the Plancherel decomposition, around π. This refines and generalizes a work of Mirodikawa. It suggests to look for such type of relations in other situations.

$\mathbb{D}(G/H)$-finite τ-spherical functions on reductive symmetric spaces

S. Souaifi [So] showed how these functions appear as linear combinations of derivatives along the complex parameter λ, of Eisenstein integrals. For K-finite functions, filtrations are introduced, whose subquotients are described in terms of induced representations. The starting point is an adaptation of ideas used by J. Franke to study spaces of automorphic forms. The use of the spectral decomposition by Langlands is replaced here by the use of the Plancherel formula. For reductive p-adic groups, and for the group case, I got similar results.

Invariant harmonic analysis on real reductive symmetric spaces

The goal is to study the H-invariant eigendistributions under $\mathbb{D}(G/H)$ on G/H and to express invariant measures on certain H-orbits in terms of these distributions (cf [D3] for the work of A.Bouaziz and P.Harinck for the group case and $G(\mathbb{C})/G(\mathbb{R})$, see also [OSe]).

Harmonic analysis on p-adic reductive symmetric spaces

For the group case, the Problems (b) and (c) of the Introduction have been solved by Harish-Chandra, up to the explicit description of the discrete series. In general, the problems are open (see [HH] for interesting structural results).

References

[An] K.Ankabout, Relations d'orthogonalité de Schur généralisées pour les espaces symétriques réductifs , *J. Funct. Anal.*, 185 (2001), 63–110.

[A] J. Arthur, A local trace formula, *Pub. Math. I.H.E.S* , 73 (1991), 5–96.

[B1] E.van den Ban, The principal series for a reductive symmetric space I, *Ann. Sc. Ec. Norm. Sup.*, 21 (1988), 359–412.

[B2] E.van den Ban, The principal series for a reductive symmetric space II, *J. of Funct. Analysis*, 109 (1992), 331–441.

[BCD] E.van den Ban, J.Carmona, P.Delorme, Paquets d'ondes dans l'espace de Schwartz d'un espace symétrique réductif , *J. of Funct. Analysis* , 139 (1996), 225–243.

[BS1] E.van den Ban , H.Schlictkrull, Fourier transform on a semisimple symmetric space, Preprint Universiteit Utrecht, No 888, Nov. 1994.

[BS2] E. van den Ban , H. Schlictkrull, The most continuous part of the Plancherel decomposition for a reductive symmetric space , *Ann. of Math.*, 145 (1997) , 267–364.

[BS3] E.van den Ban , H.Schlictkrull, Harmonic analysis on reductive symmetric spaces , *European Congress of Mathematics, Barcelona, 2000, Vol 1*, 565–582, Birkhauser Verlag, Basel, 2001.

[BS4] E.van den Ban , H.Schlictkrull, The Plancherel decomposition for a reductive symmetric space I, II, Preprints 2001.

[Be] J.N. Bernstein, On the support of the Plancherel measure, *J. Geom. Phys.*, 5 (1988), 663–710.

[BrD] J.L.Brylinski, P.Delorme, Vecteurs distributions H-invariants pour les séries principales généralisées d'espaces symétriques réductifs et prolongement méromor-
-phe d'intégrales d'Eisenstein , *Inv. Math.*, 109 (1992), 619–664.

[C1] J.Carmona, it Terme constant des fonctions tempérées sur un espace symétrique réductif,, *J. fur Reine Angew. Math.*, 491 (1997), 17–63.

[C2] J.Carmona, Plongement de séries discrètes pour un espace symétrique réductif, *J. of Funct. Analysis*, 182 (2001), 16–51.

[CD1] J.Carmona, P.Delorme, Base méromorphe de vecteurs distributions H-invariants pour les séries principales généralisées d'espaces symétriques réductifs, *J. of Funct. Analysis*, 122 (1994), 152–221.

[CD2] J.Carmona, P.Delorme, Transformation de Fourier pour les espaces symétriques réductifs , *Invent. Math.*, 134 (1998).

[D1] P.Delorme, Intégrales d'Eisenstein pour les espaces symétriques réductifs. Tempérance. Majorations. Petite matrice B, *J. of Funct. Analysis*, 136 (1996), 422–509.

[D2] P.Delorme, Troncature pour les espaces symétriques réductifs , *Acta Math.*, 179 (1997), 41–77.

[D3] P.Delorme, Inversion des intégrales orbitales sur certains espaces symétriques réductifs, d'après A. Bouaziz et P. Harinck, *Séminaire Bourbaki, Vol. 1995-1996*, Astérisque 241, Exp. 810, 417–452.

[D4] P.Delorme, Formule de Plancherel pour les espaces symétriques réductifs , *Ann. of Math.*, 147 (1998), 417–452.

[D5] P.Delorme, Espace de Schwartz pour la transformation de Fourier hypergéomé-
trique , Appendice par M. TINFOU, *J. Funct. Anal.* , 168 (1999), 239–312.

[D6] P.Delorme, The Plancherel formula on reductive symmetric spaces from the point of view of the Schwartz space. Lectures for the European Summer School of Group Theory, CIRM, Luminy 2001.

[F-J] M.Flensted-Jensen, Discrete series for semisimple symmetric spaces , *Ann. of Math.*, 111(1980), 253–311.

[Ha] P.Harinck, Fonctions orbitales sur $G_{\mathbb{C}}/G_{\mathbb{R}}$. Formule d'inversion des intégrales orbitales et formule de Plancherel , *J. Funct. Anal.*, 153 (1998), 52–107.

[H-C1] Harish-Chandra, Harmonic analysis on real reductive groups I , *J. Funct. Anal.*, 36 (1976), 1–35.

[H-C2] Harish-Chandra, Harmonic analysis on real reductive groups II, *Inventiones Math.* , 36 (1976), 1–35.

[H-C3] Harish-Chandra, Harmonic analysis on real reductive groups III. The Maass-Selberg relations and the Plancherel formula , *Ann. of Math.*, 104 (1976), 117–201.

[He] S. Helgason, *Groups and geometric analysis*, Academic Press, 1984.

[HH] A.G.Helminck, G.F. Helminck , A class of parabolic k-subgroups associated with symmetric k-varieties , *Trans. Amer. Math. Soc.*, 350 (1998), 4669–4684.

[KZ] A.Knapp, G.Zuckerman, Classification of irreducible tempered representations of semisimple groups I , *Ann. of Math.* 116 (1982), 389–455.

[Ol] G.Olafsson, Fourier and Poisson transformation associated to a semisimple symmetric space, *Inv. Math.*, 90 (1987), 1–51.

[Op] E.Opdam, Cuspidal hypergeometric functions , dedicated to R. Askey, Part I,*Methods Appl. Anal.* 6 (1999) 67–80.

[O1] T.Oshima, Fourier analysis on semisimple symmetric spaces, *Non commutative harmonic analysis and Lie groups (Marseille-Luminy, 1980)*, Lecture Notes in Math., vol 880, Springer, 1991, 357–369.

[O2] T.Oshima, Asymptotic behaviour of spherical functions on semisimple symmetric spaces , *Adv. Studies in Pure Math.* 14 (1988), 561–601.

[OM] T.Oshima, T.Matsuki, A description of discrete series for semisimple symmetric spaces, *Adv. Studies in Pure Math.*, 4 (1984), 331–390.

[OSe] T.Oshima, J.Sekiguchi, Eigenspaces of invariant differential operators on a semisimple symmetric space, *Inv. Math.* 57 (1980), 1–81.

[S] H.Schlichtkrull, The Langlands parameter of Flensted-Jensen's discrete series for semisimple symmetric spaces ,*J. of Funct Anal.*, 50 (1983), 133–150.

[So] S.Souaifi , Fonctions $\mathbb{D}(G/H)$-finies sur un espace symétrique réductif, Preprint, 2001, to appear in *J. of Funct Anal.*

[V] D. Vogan, Irreducibility of discrete series representation for semisimple symmetric spaces , *Adv. Studies in Pure Math.*, 14 (1988), 191–221.

[W] N. Wallach, *Real Reductive Groups II*, Academic Press, Inc., Boston, 1992.

ICM 2002 · Vol. II · 555–570

On the Dynamical Yang-Baxter Equation

Pavel Etingof*

(*To Mira*)

Abstract

This talk is inspired by two previous ICM talks, by V. Drinfeld (1986), and G.Felder (1994). Namely, one of the main ideas of Drinfeld's talk is that the quantum Yang-Baxter equation, which is an important equation arising in quantum field theory and statistical mechanics, is best understood within the framework of Hopf algebras, or quantum groups. On the other hand, in Felder's talk, it is explained that another important equation of mathematical physics, the star-triangle relation, may (and should) be viewed as a generalization of the quantum Yang-Baxter equation, in which solutions depend on additional "dynamical" parameters. It is also explained there that to a solution of the quantum dynamical Yang-Baxter equation one may associate a kind of quantum group. These ideas gave rise to a vibrant new branch of "quantum algebra", which may be called the theory of dynamical quantum groups. My goal in this talk is to give a bird's eye review of this theory and its applications.

The quantum dynamical Yang-Baxter equation (QDYBE) is an equation with respect to a function $R : \mathfrak{h}^* \to \mathrm{End}_{\mathfrak{h}}(V \otimes V)$, where $\mathfrak{h}$ is a commutative finite dimensional Lie algebra, and V is a semisimple $\mathfrak{h}$-module. It reads

$$R^{12}(\lambda - h^3)R^{13}(\lambda)R^{23}(\lambda - h^1) = R^{23}(\lambda)R^{13}(\lambda - h^2)R^{12}(\lambda)$$

on $V \otimes V \otimes V$, where for instance $R^{12}(\lambda - h^3)$ is defined by the formula $R^{12}(\lambda - h^3)(v_1 \otimes v_2 \otimes v_3) := \left(R^{12}(\lambda - \mu)(v_1 \otimes v_2)\right) \otimes v_3$ if v_3 is of weight μ under $\mathfrak{h}$. If $\mathfrak{h} = 0$, this equation turns into the usual quantum Yang-Baxter equation.

I will start with explaining how rational solutions of QDYBE arise already in the classical representation theory of finite dimensional simple Lie algebras, via the so called fusion construction ((the role of $\mathfrak{h}$ will be played by a Cartan subalgebra of the simple Lie algebra). Then I will explain generalizations of this construction to quantum groups, affine Lie algebras, quantum affine algebras, which yields trigonometric and elliptic solutions of QDYBE. I will then define Felder's elliptic quantum group, and formulate a q-analog of the Kazhdan-Lusztig equivalence between representations of affine Lie algebras and quantum groups.

*Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: etingof@math.mit.edu

After this I will define the classical dynamical Yang-Baxter equation (the classical limit of QDYBE), and discuss the classification result (joint with Varchenko), that says, roughly, that all solutions are of the above three types – rational, trigonometric, and elliptic. This is a dynamical version of a result of Belavin and Drinfeld.

In the second part of the talk I will consider applications of the theory of QDYBE to integrable systems. Namely, I will use the dynamical transfer matrix construction to define generalized Macdonald operators, and then construct their common eigenfunctions as (renormalized) traces of the form $\mathrm{Tr}(\Phi_\mu q^{2\lambda})$, where $\Phi_\mu : M_\mu \to M_\mu \otimes V$ is an intertwining operator between a Verma module M_μ over a quantum group tensor product with a finite dimensional module V, and with a finite dimensional module V, and $\lambda \in \mathfrak{h}^* = \mathfrak{h}$. In a special case, this yields a representation theoretic construction of type A Macdonald functions.

2000 Mathematics Subject Classification: 17B37, 20G42, 81R50.

1. Introduction

This talk is inspired by two previous ICM talks, [Dr] and [Fe]. Namely, one of the main ideas of [Dr] is that the quantum Yang-Baxter equation (QYBE), which is an important equation arising in quantum field theory and statistical mechanics, is best understood within the framework of Hopf algebras, or quantum groups. On the other hand, in [Fe], it is explained that another important equation of mathematical physics, the star-triangle relation, may (and should) be viewed as a generalization of QYBE, in which solutions depend on additional "dynamical" parameters. It is also explained there that to a solution of the quantum dynamical Yang-Baxter equation one may associate a kind of quantum group. These ideas gave rise to a vibrant new branch of "quantum algebra", which may be called the theory of dynamical quantum groups. My goal in this talk is to give a bird's eye review of some aspects of this theory and its applications.

Remark Because of size restrictions, the list of references below is not complete. The reader is referred to original papers and review [ES2] for more references.

Acknowledgements I am indebted to G. Felder for creating this subject, and deeply grateful to I. Frenkel, A. Kirillov Jr., A. Moura, O.Schiffmann, and A.Varchenko for collaboration that led to this work. This work was partially supported by the NSF grant DMS-9988796, and was done in part for the Clay Mathematics Institute. I am grateful to the Max Planck Institute for hospitality.

2. The quantum dynamical Yang-Baxter equation

2.1. The equation

The quantum dynamical Yang-Baxter equation (QDYBE), proposed by Gervais, Neveu, and Felder, is an equation with respect to a (meromorphic) function $R : \mathfrak{h}^* \to \mathrm{End}_{\mathfrak{h}}(V \otimes V)$, where $\mathfrak{h}$ is a commutative finite dimensional Lie algebra over $\mathbb{C}$, and V is a semisimple $\mathfrak{h}$-module. It reads

$$R^{12}(\lambda - h^3)R^{13}(\lambda)R^{23}(\lambda - h^1) = R^{23}(\lambda)R^{13}(\lambda - h^2)R^{12}(\lambda)$$

on $V \otimes V \otimes V$. Here h^i is the *dynamical notation*, to be extensively used below: for instance, $R^{12}(\lambda - h^3)$ is defined by the formula $R^{12}(\lambda - h^3)(v_1 \otimes v_2 \otimes v_3) := \big(R^{12}(\lambda - \mu)(v_1 \otimes v_2)\big) \otimes v_3$ if v_3 is of weight μ under $\mathfrak{h}$.

It is also useful to consider QDYBE with spectral parameter. In this case the unknown function R depends meromorphically on an additional complex variable u. Let $u_{ij} = u_i - u_j$. Then the equation reads

$$R^{12}(u_{12}, \lambda - h^3)R^{13}(u_{13}, \lambda)R^{23}(u_{23}, \lambda - h^1) = R^{23}(u_{23}, \lambda)R^{13}(u_{13}, \lambda - h^2)R^{12}(u_{12}, \lambda).$$

Sometimes it is necessary to consider QDYBE with step $\gamma \in \mathbb{C}^*$, which differs from the usual QDYBE by the replacement $h^i \to \gamma h^i$. Clearly, $R(\lambda)$ satisfies QDYBE iff $R(\lambda/\gamma)$ satisfies QDYBE with step γ.

Invertible solutions of QDYBE are called quantum dynamical R-matrices (with or without spectral parameter, and with step γ if needed). If $\mathfrak{h} = 0$, QDYBE turns into the usual quantum Yang-Baxter equation $R^{12}R^{13}R^{23} = R^{23}R^{13}R^{12}$.

2.2. Examples of solutions of QDYBE

Let V be the vector representation of $sl(n)$, and $\mathfrak{h}$ the Lie algebra of traceless diagonal matrices. In this case $\lambda \in \mathfrak{h}^*$ can be written as $\lambda = (\lambda_1, ..., \lambda_n)$, where $\lambda_i \in \mathbb{C}$. Let v_a, $a = 1, ..., n$ be the standard basis of V. Let E_{ab} be the matrix units given by $E_{ab}v_c = \delta_{bc}v_a$.

We will now give a few examples of quantum dynamical R-matrices. The general form of the R-matrices will be

$$R = \sum_a E_{aa} \otimes E_{aa} + \sum_{a \neq b} \alpha_{ab} E_{aa} \otimes E_{bb} + \sum_{a \neq b} \beta_{ab} E_{ab} \otimes E_{ba}, \tag{1}$$

where α_{ab}, β_{ab} are functions which will be given explicitly in each example.

Example 1 The basic rational solution. Let $\beta_{ab} = \frac{1}{\lambda_b - \lambda_a}$, $\alpha_{ab} = 1 + \beta_{ab}$. Then $R(\lambda)$ is a dynamical R-matrix (with step 1).

Example 2 The basic trigonometric solution. Let $\beta_{ab} = \frac{q-1}{q^{\lambda_b - \lambda_a} - 1}$, $\alpha_{ab} = q + \beta_{ab}$. Then $R(\lambda)$ is a dynamical R-matrix (also with step 1).

Example 3 The basic elliptic solution with spectral parameter (Felder's solution). Let $\theta(u, \tau) = -\sum_{j \in \mathbb{Z}+1/2} e^{\pi i(j^2\tau + 2j(u+1/2))}$ be the standard theta-function (for brevity will write it as $\theta(u)$). Let $\beta_{ab} = \frac{\theta(u - \lambda_b + \lambda_a)\theta(\gamma)}{\theta(u-\gamma)\theta(\lambda_b - \lambda_a)}$, $\alpha_{ab} = \frac{\theta(\lambda_a - \lambda_b + \gamma)\theta(u)}{\theta(\lambda_a - \lambda_b)\theta(u-\gamma)}$. These functions can be viewed as functions of $z = e^{2\pi i u}$. Then $R(u, \lambda)$ is a quantum dynamical R-matrix with spectral parameter and step γ. One may also define the basic trigonmetric and basic rational solution with spectral parameter, which differ from the elliptic solution by replacement of $\theta(x)$ by $\sin(x)$ and x, respectively.

Remark 1 In examples 1 and 2, the dynamical R-matrix satisfies the *Hecke condition* $(PR-1)(PR+q)=0$, with $q=1$ in example 1 (where P is the permutation on $V\otimes V$), and in example 3 the unitarity condition $R(u,\lambda)R^{21}(-u,\lambda)=1$.

Remark 2 The basic trigonometric solution degenerates into the basic rational solution as $q\to 1$. Also, the basic elliptic solution with spectral parameter can be degenerated into the basic trigonometric and rational solutions with spectral parameter by renormalizing variables and sending one, respectively both periods of theta functions to infinity (see [EV2]). Furthermore, the basic trigonometric and rational solutions with spectral parameter can, in turn, be degenerated into the solutions of Examples 1,2, by sending the spectral parameter to infinity. Thus, in essense, all the examples we gave are obtained from Felder's solution.

2.3. The tensor category of representations associated to a quantum dynamical R-matrix

Let R be a quantum dynamical R-matrix with spectral parameter. According to [Fe], a representation of R is a semisimple $\mathfrak{h}$-module W and an invertible meromorphic function $L=L_W:\mathbb{C}\times\mathfrak{h}^*\to \mathrm{End}_{\mathfrak{h}}(V\otimes W)$, such that

$$R^{12}(u_{12},\lambda-h^3)L^{13}(u_{13},\lambda)L^{23}(u_{23},\lambda-h^1)=L^{23}(u_{23},\lambda)L^{13}(u_{13},\lambda-h^2)R^{12}(u_{12},\lambda). \tag{2}$$

(In the case of step γ, h^i should be replaced by γh^i).

For example: $(\mathbb{C},1)$ (trivial representation) and (V,R) (vector representation).

A morphism $f:(W,L_W)\to(W',L_{W'})$ is a meromorphic function $f:\mathfrak{h}^*\to \mathrm{End}_{\mathfrak{h}}(W)$ such that $(1\otimes f(\lambda))L_W(u,\lambda)=L_W(u,\lambda)(1\otimes f(\lambda-\gamma h^1))$. With this definition, representations form an (additive) category $\mathrm{Rep}(R)$. Moreover, it is a tensor category [Fe]: given (W,L_W) and (U,L_U), one can form the tensor product representation $(W\otimes U,L_{W\otimes U})$, where $L_{W\otimes U}(u,\lambda)=L_W^{12}(u,\lambda-\gamma h^3)L_U^{13}(u,\lambda)$; tensor product of morphisms is defined by $(f\otimes g)(\lambda)=f(\lambda-\gamma h^2)\otimes g(\lambda)$.

In absence of spectral parameter, one should use the same definitions without u.

2.4. Gauge transformations and classification

There exists a group of rather trivial transformations acting on quantum dynamical R-matrices with step γ. They are called *gauge transformations.* If $\mathfrak{h}$ and V are as in the previous section, then gauge transformations (for dynamical R-matrices without spectral parameter) are:

1. Twist by a closed multiplicative 2-form ϕ: $\alpha_{ab}\to\alpha_{ab}\phi_{ab}$, where $\phi_{ab}=\phi_{ba}^{-1}$, and $\phi_{ab}(\lambda)\phi_{bc}(\lambda)\phi_{ca}(\lambda)=\phi_{ab}(\lambda-\gamma\omega_c)\phi_{bc}(\lambda-\gamma\omega_a)\phi_{ca}(\lambda-\gamma\omega_b)$ $(\omega_i=\mathrm{weight}(v_i))$.

2. Permutation of indices $a=1,...,n$; $\lambda\to\lambda-\nu$; $R\to cR$.

In presence of the spectral parameter u, the constant in c in 2 may depend on u, and there are the following additional gauge transformations:

3. Given meromorphic $\psi:\mathfrak{h}^*\to\mathbb{C}$, $\alpha_{ab}\to e^{u(\psi(\lambda)-2\psi(\lambda-\gamma\omega_a)+\psi(\lambda-\gamma\omega_a-\gamma\omega_b))}\alpha_{ab}$, $\beta_{ab}\to e^{u(\psi(\lambda)-\psi(\lambda-\gamma\omega_a)-\psi(\lambda-\gamma\omega_b)+\psi(\lambda-\gamma\omega_a-\gamma\omega_b))}\beta_{ab}$.

4. Multiplication of u by a constant.

Theorem 2.1 [EV2] *Any quantum dynamical R-matrix for* $\mathfrak{h}, V$ *satisfying the Hecke condition with* $q = 1$ *(respectively,* $q \neq 1$*) is a gauge transformation of the basic rational (respectively, trigonometric) solution, or a limit of such R-matrices.*

In the spectral parameter case, a similar result is known only under rather serious restrictions (see [EV2], Theorem 2.5). For more complicated configurations $(\mathfrak{h}, V)$, classification is not available.

Remark Gauge transformations 2-4 do not affect the representation category of the R-matrix. Gauge transformation 1 does not affect the category if the closed form ϕ is exact: $\phi = d\xi$, i.e. $\phi_{ab}(\lambda) = \xi_a(\lambda)\xi_b(\lambda - \gamma\omega_a)\xi_a(\lambda - \gamma\omega_b)^{-1}\xi_b(\lambda)^{-1}$, where $\xi_a(\lambda)$ is a collection of meromorphic functions. We note that this is a very mild condition: for example, if γ is a formal paremeter and we work with analytic functions of λ in a simply connected domain, then a multiplicative 2-form is closed iff it is exact ("multiplicative Poincare lemma").

2.5. Dynamical quantum groups

Equation 2 may be regarded as a set of defining relations for an associative algebra A_R (see [EV2] for precise definitions). This algebra is a dynamical analogue of the quantum group attached to an R-matrix defined in [FRT], and representations of R are an appropriate class of representations of this algebra. The algebra A_R is called the dynamical quantum group attached to R. If R is the basic elliptic solution then A_R is the elliptic quantum group of [Fe]. The structure and representations of A_R are studied in many papers (e.g. [Fe, EV2, FV1, TV]).

To keep this paper within bounds, we will not discuss A_R in detail. However, let us mention ([EV2]) that A_R is a bialgebroid with base $\mathfrak{h}^*$. This corresponds to the fact that the category $\mathrm{Rep}(R)$ is a tensor category. Moreover, if R satisfies an additional rigidity assumption (valid for example for the basic rational and trigonometric solutions) then the category $\mathrm{Rep}(R)$ has duality, and A_R is a Hopf algebroid, or a quantum groupoid (i.e. it has an antipode).

Remark For a general theory of bialgebroids and Hopf algebroids the reader is referred to [Lu]. However, let us mention that bialgebroids with base X correspond to pairs (tensor category, tensor functor to $O(X)$-bimodules), similarly to how bialgebras correspond to pairs (tensor category, tensor functor to vector spaces) (i.e. via Tannakian formalism).

2.6. The classical dynamical Yang-Baxter equation

Recall that if $R = 1 - \hbar r + O(\hbar^2)$ is a solution QYBE, then r satisfies the classical Yang-Baxter equation (CYBE), $[r^{12}, r^{13}] + [r^{12}, r^{23}] + [r^{13}, r^{23}] = 0$, and that r is called the classical limit of R, while R is called a quantization of r. Similarly, let $R(\lambda, \hbar)$ be a family of solutions of QDYBE with step $\hbar$ given by a series $1 - \hbar r(\lambda) + O(\hbar^2)$. Then it is easy to show that $r(\lambda)$ satisfies the following differential equation, called the *classical dynamical Yang-Baxter equation* (CDYBE):

$$\sum_i \left(x_i^{(1)} \frac{\partial r^{23}}{\partial x_i} - x_i^{(2)} \frac{\partial r^{13}}{\partial x_i} + x_i^{(3)} \frac{\partial r^{12}}{\partial x_i} \right) + [r^{12}, r^{13}] + [r^{12}, r^{23}] + [r^{13}, r^{23}] = 0, \quad (3)$$

(where x_i is a basis of $\mathfrak{h}$). The function $r(\lambda)$ is called the classical limit of $R(\lambda,\hbar)$, and $R(\lambda,\hbar)$ is called a quantization of $r(\lambda)$.

Define a classical dynamical r-matrix to be a meromorphic function $r:\mathfrak{h}^*\to \mathrm{End}_{\mathfrak{h}}(V\otimes V)$ satisfying CDYBE.

Conjecture 2.2 *Any classical dynamical r-matrix can be quantized.*

This conjecture is known in the non-dynamical case ([EK]), and was proved in [Xu] in the dynamical case for skew-symmetric solutions ($r^{21}=-r$) satisfying additional technical assumptions. A general proof for the skew-symmetric case is recently proposed in a preprint by T.Mochizuki. However, the most interesting non-skew-symmetric case is still open.

Similarly, the classical limit of QDYBE with spectral parameter is CDYBE with spectral parameter. It is an equation with respect to $r(u,\lambda)$ and differs from the usual CDYBE by insertion of u_{ij} as an additional argument of r^{ij}.

Remark 1 Similarly to CYBE, CDYBE makes sense for functions with values in $\mathfrak{g}\otimes\mathfrak{g}$, where $\mathfrak{g}$ is a Lie algebra containing $\mathfrak{h}$.

Remark 2 The classical limit of the notion of a quantum groupoid is the notion of a Poisson groupoid, due to Weinstein. By definition, a Poisson groupoid is a groupoid G which is also a Poisson manifold, such that the graph of the multiplication is coisotropic in $G\times G\times \bar{G}$ (where $\bar{G}$ is G with reversed sign of Poisson bracket). Such a groupoid can be attached ([EV1]) to a classical dynamical r-matrix $r:\mathfrak{h}^*\to\mathfrak{g}\otimes\mathfrak{g}$, such that $r^{21}+r$ is constant and invariant (i.e. r is a "dynamical quasitriangular structure" on $\mathfrak{g}$). This is the classical limit of the assignment of a quantum groupoid to a quantum dynamical R-matrix ([EV2]).

2.7. Examples of solutions of CDYBE

We will give examples of solutions of CDYBE in the case when $\mathfrak{g}$ is a finite dimensional simple Lie algebra, and $\mathfrak{h}$ is its Cartan subalgebra. We fix an invariant inner product on $\mathfrak{g}$, It restricts to a nondegenerate inner product on $\mathfrak{h}$. Using this inner product, we identify $\mathfrak{h}^*$ with $\mathfrak{h}$ ($\lambda\in\mathfrak{h}^*\to\bar{\lambda}\in\mathfrak{h}$), which yields an inner product on $\mathfrak{h}^*$. The normalization of the inner product is chosen so that short roots have squared length 2. Let x_i be an orthonormal basis of $\mathfrak{h}$,and let e_α, $e_{-\alpha}$ denote positive (respectively, negative) root elements of $\mathfrak{g}$, such that $(e_\alpha,e_{-\alpha})=1$.

Example 1 The basic rational solution $r(\lambda)=\sum_{\alpha>0}\frac{e_\alpha\wedge e_{-\alpha}}{(\lambda,\alpha)}$.

Example 2 The basic trigonometric solution $r(\lambda)=\frac{\Omega}{2}+\sum_{\alpha>0}\frac{1}{2}\mathrm{cotanh}(\frac{(\lambda,\alpha)}{2})$, where $\Omega\in S^2\mathfrak{g}$ is the inverse element to the inner product on $\mathfrak{g}$.

Example 3 The basic elliptic solution with spectral parameter (Felder's solution) $r(u,\lambda)=\frac{\theta'(u)}{\theta(u)}\sum_i x_i\otimes x_i+\sum_\alpha\frac{\theta(u+(\lambda,\alpha))\theta'(0)}{\theta((\lambda,\alpha))\theta(u)}e_\alpha\otimes e_{-\alpha}$.

Remark 1 One says that a classical dynamical r-matrix r has coupling constant ε if $r+r^{21}=\varepsilon\Omega$. If r is with spectral parameter, one says that it has coupling constant ε if $r(u,\lambda)+r^{21}(-u,\lambda)=0$, and $r(u,\lambda)$ has a simple pole at $u=0$ with residue $\varepsilon\Omega$ (these are analogs of the Hecke and unitarity conditions in the quantum case). With these definitions, the basic rational solution has coupling constant 0, while the trigonometric and elliptic solutions have coupling constant 1.

Remark 2 As in the quantum case, Example 3 can be degenerated into its trigonometric and rational versions where $\theta(x)$ is replaced by $\sin(x)$ and x, respectively, and Examples 1 and 2 can be obtained from Example 3 by a limit.

Remark 3 The classical limit of the basic rational and trigonometric solutions of QDYBE (modified by $\lambda \to \lambda/\hbar$) is the basic rational, respectively trigoinometric, solutions of CDYBE for $\mathfrak{g} = \mathfrak{sl}_n$ (in the trigonometric case we should set $q = e^{-\hbar/2}$). The same is true for the basic elliptic solution (with $\gamma = \hbar$).

Remark 4 These examples make sense for any reductive Lie algebra $\mathfrak{g}$.

2.8. Gauge transformations and classification of solutions for CDYBE

It is clear from the above that it is interesting to classify solutions of CDYBE. As in the quantum case, it should be done up to gauge transformations. These transformations are classical analogs of the gauge transformations in the quantum case, and look as follows.

1. $r \to r+\omega$, where $\omega = \sum_{i,j} C_{ij}(\lambda) x_i \wedge x_j$ is a meromorphic closed differential 2-form on $\mathfrak{h}^*$.

2. $r(u,\lambda) \to ar(a\lambda - \nu)$; Weyl group action.

In the case of spectral parameter, there are additional transformations:

3. $u \to bu$.

4. Let $r = \sum_{i,j} S_{ij} x_i \otimes x_j + \sum_\alpha \phi_\alpha e_\alpha \otimes e_{-\alpha}$. The transformation is $S_{ij} \to S_{ij} + u\frac{\partial^2 \psi}{\partial x_i \partial x_j}$, $\phi_\alpha \to \phi_\alpha e^{u\partial_\alpha \psi}$, where ψ is a function on $\mathfrak{h}^*$ with meromorphic $d\psi$.

Theorem 2.3 [EV1] *(i) Any classical dynamical r-matrix with zero coupling constant is a gauge transformation of the basic rational solution for a reductive subalgebra of $\mathfrak{g}$ containing $\mathfrak{h}$, or its limiting case.*

(ii) Any classical dynamical r-matrix with nonzero coupling constant is a gauge transformation of the basic trigonometric solution for $\mathfrak{g}$, or its limiting case.

(iii) Any classical dynamical r-matrix with spectral parameter and nonzero coupling constant is a gauge transformation of the basic elliptic solution for $\mathfrak{g}$, or its limiting case.

Remark One may also classify dynamical r-matrices with nonzero coupling constant (without spectral parameter) defined on $\mathfrak{l}^*$ for a Lie subalgebra $\mathfrak{l} \subset \mathfrak{h}$, on which the inner product is nondegenerate ([Sch]). Up to gauge transformations they are classified by generalized Belavin-Drinfeld triples, i.e. triples (Γ_1, Γ_2, T), where Γ_i are subdiagrams of the Dynkin diagram Γ of $\mathfrak{g}$, and $T : \Gamma_1 \to \Gamma_2$ is a bijection perserving the inner product of simple roots (so this classification is a dynamical analog of the Belavin-Drinfeld classification of r-matrices on simple Lie algebras, and the classification of [EV1] is the special case $\Gamma_1 = \Gamma_2 = \Gamma$, $T = 1$). The formula for a classical dynamical r-matrix corresponding to such a triple given in [Sch] works for any Kac-Moody algebra, and in the case of affine Lie algebras yields classical dynamical r-matrices with spectral parameter (however, the classification is this case has not been worked out). Explicit quantization of the dynamical r-matrices from [Sch] (for any Kac-Moody algebra) is given in [ESS].

3. The fusion and exchange construction

Unlike QYBE, interesting solutions of QDYBE may be obtained already from classical representation theory of Lie algebras. This can be done through the fusion and exchange construction, see [Fa, EV3].

3.1. Intertwining operators

Let $\mathfrak{g}$ be a simple finite dimensional Lie algebra over $\mathbb{C}$, with polar decomposition $\mathfrak{g} = \mathfrak{n}_+ \oplus \mathfrak{h} \oplus \mathfrak{n}_-$. For any $\mathfrak{g}$-module V, we write $V[\nu]$ for the weight subspace of V of weight $\nu \in \mathfrak{h}^*$. Let M_λ denote the Verma module over $\mathfrak{g}$ with highest weight $\lambda \in \mathfrak{h}^*$, x_λ its highest weight vector, and x_λ^* the lowest weight vector of the dual module. Let V be a finite dimensional representation of $\mathfrak{g}$. Consider an intertwining operator $\Phi : M_\lambda \to M_\mu \otimes V$. The vector $x_\mu^*(\Phi x_\lambda) \in V[\lambda - \mu]$ is called the expectation value of Φ, and denoted $\langle \Phi \rangle$.

Lemma 3.1 *If M_μ is irreducible (i.e. for generic μ), the map $\Phi \to \langle \Phi \rangle$ is an isomorphism* $\mathrm{Hom}_\mathfrak{g}(M_{\mu+\nu}, M_\mu \otimes V) \to V[\nu]$.

Lemma 3.1 allows one to define for any $v \in V[\nu]$ (and generic λ) the intertwining operator $\Phi_\lambda^v : M_\lambda \to M_{\lambda-\nu} \otimes V$, such that $\langle \Phi_\lambda^v \rangle = v$.

3.2. The fusion and exchange operators

Now let V, W be finite dimensional $\mathfrak{g}$-modules, and $v \in V, w \in W$ homogeneous vectors, of weights $\mathrm{wt}(v), \mathrm{wt}(w)$. Consider the composition of two intertwining operators

$$\Phi_\lambda^{w,v} := (\Phi^w_{\lambda-\mathrm{wt}(v)} \otimes 1)\Phi_\lambda^v : M_\lambda \to M_{\lambda-\mathrm{wt}(v)-\mathrm{wt}(w)} \otimes W \otimes V.$$

The expectation value of this composition, $\langle \Phi_\lambda^{w,v} \rangle$, is a bilinear function of w and v. Therefore, there exists a linear operator $J_{WV}(\lambda) \in \mathrm{End}(W \otimes V)$ (of weight zero, i.e. commuting with $\mathfrak{h}$), such that $\langle \Phi_\lambda^{w,v} \rangle = J_{WV}(\lambda)(w \otimes v)$. In other words, we have $(\Phi^w_{\lambda-\mathrm{wt}(v)} \otimes 1)\Phi_\lambda^v = \Phi_\lambda^{J_{WV}(\lambda)(w\otimes v)}$. The operator $J_{WV}(\lambda)$ is called the *fusion operator* (because it tells us how to "fuse" two intertwining operators).

The fusion operator has a number of interesting properties, which we discuss below. In particular, it is lower triangular, i.e. has the form $J = 1 + N$, where N is a sum of terms which have strictly positive weights in the second component. Consequently, N is nilpotent, and J is invertible.

Define also the *exchange operator* $R_{VW}(\lambda) := J_{VW}^{-1}(\lambda) J_{WV}^{21}(\lambda) : V \otimes W \to V \otimes W$. This operator tells us how to exchange the order of two intertwining operators, in the sense that if $R_{WV}(\lambda)(w \otimes v) = \sum_i w_i \otimes v_i$ (where w_i, v_i are homogeneous), then $\Phi_\lambda^{w,v} = P \sum_i \Phi_\lambda^{v_i,w_i}$ (where P permutes V and W).

Theorem 3.2 [EV3] *$R_{VV}(\lambda)$ is a solution of QDYBE.*

3.3. Fusion and exchange operators for quantum groups

The fusion and exchange constructions generalize without significant changes to the case when the Lie algebra $\mathfrak{g}$ is replaced by the quantum group $U_q(\mathfrak{g})$, where q is not zero or a root of unity. The only change that needs to be made is in the definition of the exchange operator: namely, one sets $R(\lambda) = J_{VW}^{-1}(\lambda)\mathcal{R}^{21}J_{WV}^{21}(\lambda)$, where $\mathcal{R}$ is the universal R-matrix of $U_q(\mathfrak{g})$. This is because when changing the order of intertwining operators, we must change the order of tensor product of representations $V \otimes W$, which in the quantum case is done by means of the R-matrix.

Example 3.3 *Let $\mathfrak{g} = \mathfrak{sl}_n$, and V be the vector representation of $U_q(\mathfrak{g})$. Then the exchange operator has the form $R = q^{1-1/n}\tilde{R}$, where $\tilde{R}$ is given by (1), with $\beta_{ab} = \frac{q^{-2}-1}{q^{2(\lambda_a-\lambda_b-a+b)}-1}$, $\alpha_{ab} = q^{-1}$ for $a < b$, and $\alpha_{ab} = \frac{(q^{2(\lambda_b-\lambda_a+a-b)}-q^{-2})(q^{2(\lambda_b-\lambda_a+a-b)}-q^2)}{q(q^{2(\lambda_b-\lambda_a+a-b)}-1)^2}$ if $a > b$. The exchange operator for the vector representation of $\mathfrak{g}$ is obtained by passing to the limit $q \to 1$; i.e. it is given by (1), with $\beta_{ab} = \frac{1}{\lambda_b-\lambda_a-b+a}$, $\alpha_{ab} = 1$ for $a < b$, and $\alpha_{ab} = \frac{(\lambda_b-\lambda_a+a-b-1)(\lambda_b-\lambda_a+a-b+1)}{(\lambda_b-\lambda_a+a-b)^2}$ if $a > b$. It is easy to see that these exchange operators are gauge equivalent to the basic rational and trigonometric solutions of QDYBE, respectively.*

3.4. The ABRR equation

The fusion operator is not only a tool to define the exchange operator satisfying QDYBE, but is an interesting object by itself, which deserves a separate study; so we will briefly discuss its properties.

Let ρ the half-sum of positive roots. Let $\Theta(\lambda) \in U(\mathfrak{h})$ be given by $\Theta(\lambda) = \bar{\lambda} + \bar{\rho} - \frac{1}{2}\sum x_i^2$. Then $\Theta(\lambda)$ defines an operator in any $U_q(\mathfrak{g})$-module with weight decomposition. Let $\mathcal{R}_0 = \mathcal{R}q^{-\sum x_i \otimes x_i}$ be the unipotent part of the universal R-matrix.

Theorem 3.4 (ABRR equation, [ABRR, JKOS]) *For $q \neq 1$, the fusion operator is a unique lower triangular zero weight operator, which satisfies the equation:*

$$J(\lambda)(1 \otimes q^{2\Theta(\lambda)}) = \mathcal{R}_0^{21}(1 \otimes q^{2\Theta(\lambda)})J(\lambda). \tag{4}$$

For $q = 1$, the fusion operator satisfies the classical limit of this equation:

$$[J(\lambda), 1 \otimes \Theta(\lambda)] = (\sum_{\alpha>0} e_{-\alpha} \otimes e_\alpha)J(\lambda), \tag{5}$$

(Here for brevity we have dropped the subscripts W and V, with the understanding that both sides are operators on $W \otimes V$).

3.5. The universal fusion operator

Using the ABRR equation, we can define the universal fusion operator, living in a completion of $U_q(\mathfrak{g})^{\otimes 2}$, which becomes $J_{WV}(\lambda)$ after evaluating in $W \otimes V$.

Namely, the universal fusion operator $J(\lambda)$ is the unique lower triangular solution of the ABRR equation in a completion of $U_q(\mathfrak{g})^{\otimes 2}$. This solution can be found in the form of a series $J = \sum_{n\geq 0} J_n$, $J_0 = 1$, where $J_n \in U_q(\mathfrak{g}) \otimes U_q(\mathfrak{g})$ has zero weight and its second component has degree n in principal gradation; so J_n are computed recursively.

This allows one to compute the universal fusion operator quite explicitly. For example, if $q = 1$ and $\mathfrak{g} = \mathfrak{sl}_2$, then the universal fusion operator is given by the formula $J(\lambda) = \sum_{n\geq 0} \frac{(-1)^n}{n!} f^n \otimes (\lambda - h + n + 1)^{-1} \dots (\lambda - h + 2n)^{-1} e^n$.

3.6. The dynamical twist equation

Another important property of the fusion operator is the dynamical twist equation (which is a dynamical analog of the equation for a Drinfeld twist in a Hopf algebra).

Theorem 3.5 *The universal fusion operator $J(\lambda)$ satisfies the dynamical twist equation $J^{12,3}(\lambda)J^{1,2}(\lambda - h^{(3)}) = J^{1,23}(\lambda)J^{2,3}(\lambda)$.*

Here the superscripts of J stand for components where the first and second component of J acts; for example $J^{1,23}$ means $(1 \otimes \Delta)(J)$, and $J^{1,2}$ means $J \otimes 1$.

3.7. The fusion operator for affine algebras

The fusion and exchange construction can be generalized to the case when the simple Lie algebra $\mathfrak{g}$ is replaced by any Kac-Moody Lie algebra. This generalization is especially interesting if $\mathfrak{g}$ is replaced with the affine Lie algebra $\hat{\mathfrak{g}}$, and V, W are finite dimensional representations of $U_q(\hat{\mathfrak{g}})$ (where q is allowed to be 1). In this case, for each $z \in \mathbb{C}^*$ we have an outer automorphism $D_z : U_q(\hat{\mathfrak{g}}) \to U_q(\hat{\mathfrak{g}})$, which preserves the Chevalley generators q^h and e_i, f_i, $i > 0$, of $U_q(\hat{\mathfrak{g}})$, while multiplying e_0 by z and f_0 by z^{-1}. Define the shifted representation $V(z)$ by $\pi_{V(z)}(a) = \pi_V(D_z(a))$. Let $\hat{\lambda} = (\lambda, k)$ be a weight for $\hat{\mathfrak{g}}$ (k is the level). Then similarly to the finite dimensional case one can define the intertwining operator $\Phi^v_{\lambda,k}(z) : M_{\hat{\lambda}} \to M_{\hat{\lambda}-\mathrm{wt}(v)} \hat{\otimes} V(z)$ (to a completed tensor product); it can be written as an infinite series $\sum_{n\in\mathbb{Z}} \Phi^v_{\lambda,k}[n] z^{-n}$, where $\Phi^v_{\lambda,k}[n] : M_{\hat{\lambda}} \to M_{\hat{\lambda}-\mathrm{wt}(v)} \otimes V$ are linear operators. Furthermore, the expectation value of the composition $\langle (\Phi^w_{\lambda-\mathrm{wt}(v),k}(z_1) \otimes 1)\Phi^v_{\lambda,k}(z_2)\rangle$ is defined as a Taylor series in z_2/z_1. Thus, one can define the fusion operator $J_{WV}(z) \in \mathrm{End}_{\mathfrak{h}}(W \otimes V)[[z]]$ such that this expectation value is equal to $J_{WV}(z)(w \otimes v)$.

3.8. Fusion operator and correlators in the WZW model

One may show (see [FR, EFK]) that the series $J_{WV}(z)$ is convergent in some neighborhood of 0 to a holomorphic function. This function has a physical interpretation. Namely, if $q = 1$, the operators $\Phi^v_{\lambda,k}(z)$ are, essentially, vertex operators (primary fields) for the Wess-Zumino-Witten conformal field theory, and the function $J_{WV}(z)(w \otimes v)$ is a 2-point correlation function of vertex operators. If $q \neq 1$,

this function is a 2-point correlation function of q-vertex operators, and has a similar interpretation in terms of statistical mechanics.

3.9. The ABRR equation in the affine case as the KZ (q-KZ) equation

The ABRR equation has a straightforward generalization to the affine case, which also has a physical interpretation. Namely, in the case $q = 1$ it coincides with the (trigonometric) Knizhnik-Zamolodchikov (KZ) equation for the 2-point correlation function, while for $q \neq 1$ it coincides with the quantized KZ equation for the 2-point function of q-vertex operators, derived in [FR] (see also [EFK]).

One may also define (for any Kac-Moody algebra) the *multicomponent universal fusion operator* $J^{1\dots N}(\lambda) = J^{1,2\dots N}(\lambda)\dots J^{2,3\dots N}(\lambda)J^{N-1,N}(\lambda)$. It satisfies a multicomponent version of the ABRR equation. In the affine case, $J^{1\dots N}(\lambda)$ is interpreted as the N-point correlation function for vertex (respectively, q-vertex) operators, while the multicomponent ABRR equation is interpreted as the KZ (respectively, qKZ) equation for this function. See [EV5] for details.

3.10. The exchange operator for affine algebras and monodromy of KZ (q-KZ) equations

The generalization of the exchange operator to the affine case is rather tricky, and there is a serious difference between the classical ($q = 1$), and quantum ($q \neq 1$) case. The naive definition would be $R_{VW}(u, \hat{\lambda}) = J_{VW}(z, \hat{\lambda})^{-1}\mathcal{R}^{21}|_{V\otimes W(z)} J^{21}_{WV}(z^{-1}, \hat{\lambda})$ (where $z = e^{2\pi i u}$). However, this definition does not immediately make sense, since we are multiplying infinite Taylor series in z by infinite Taylor series in z^{-1}. To make sense of such product, consider the cases $q = 1$ and $q \neq 1$ separately.

If $q = 1$, the functions $J_{VW}(z, \hat{\lambda})$ and $J^{21}_{WV}(z^{-1}, \hat{\lambda})$ are both solutions of the Knizhnik-Zamolodchikov differential equation, one regular near 0 and the other near ∞. The equation is nonsingular outside $0, \infty$, and 1. Thus, the series $J_{VW}(z, \hat{\lambda})$ is convergent for $|z| < 1$. On the other hand, for $0 < z < 1$ define the function $A^{\pm}(J^{21}_{WV}(z^{-1}, \hat{\lambda}))$ to be the analytic continuation of $J^{21}_{WV}(z^{-1}, \hat{\lambda})$ from the region $z > 1$ along a curve passing 1 from above (for +) and below (for -). Then the function $R^{\pm}_{VW}(u, \hat{\lambda}) := J_{VW}(z, \hat{\lambda})^{-1}A^{\pm}(J^{21}_{WV}(z^{-1}, \hat{\lambda}))$ is of zero weight, and satisfies the QDYBE with spectral parameter (for $V = W$). The operator $R^{\pm}_{VW}(u, \hat{\lambda})$ is the appropriate analog of the exchange operator (depending on a choice of sign).

If $q \neq 1$, the functions $J_{VW}(z, \hat{\lambda})$ and $\mathcal{R}^{21}|_{V\otimes W(z)} J^{21}_{WV}(z^{-1}, \hat{\lambda})$ are both solutions of the quantized Knizhnik-Zamolodchikov equation, which is a difference equation with multiplicative step $p = q^{-2m(k+g)}$, where m is the ratio of squared norms of long and short roots, k the level of $\hat{\lambda}$, and g the dual Coxeter number of $\mathfrak{g}$. Therefore, if $|p| \neq 1$, these functions admit a meromorphic continuation to the whole $\mathbb{C}^*$ (unlike the $q = 1$ case, they are now single-valued), and the naive formula for $R_{VW}(u, \hat{\lambda})$ makes sense. As in the $q = 1$ case, this function is of zero weight and

satisfies the QDYBE with spectral parameter (for $V = W$); it is the appropriate generalization of the exchange operator (see [EFK] for details).

The operators $R^{\pm}_{VW}(z, \hat{\lambda})$, essentially, represent the monodromy of the KZ differential equation. In particular, $R^{\pm}_{VW}$ is "almost constant" in u: its matrix elements in a homogeneous basis under $\mathfrak{h}$ are powers of $e^{2\pi i u}$ (in fact, $R^{\pm}_{VV}(u, \hat{\lambda})$ is gauge equivalent, in an appropriate sense, to a solution of QDYBE without spectral parameter). Similarly, the operator $R_{VW}(z, \hat{\lambda})$ for $q \neq 1$ represents the q-monodromy (Birkhoff's connection matrix) of the q-KZ equation. In particular, the matrix elements of $R_{VW}(u, \hat{\lambda})$ are quasiperiodic in u with period $\tau = \log p / 2\pi i$. Since they are also periodic with period 1 and meromorphic, they can be expressed rationally via elliptic theta-functions.

Example 3.6 *Let $\mathfrak{g} = \mathfrak{sl}_n$, and V the vector representation of $U_q(\hat{\mathfrak{g}})$. If $q \neq 1$, the exchange operator $R_{VV}(u, \hat{\lambda})$ is a solution of QDYBE, gauge equivalent to the basic elliptic solution with spectral parameter (see [Mo] and references therein, and also [FR, EFK]). The gauge transformation involves an exact multiplicative 2-form, which expresses via q-Gamma functions with* q $= p$. *Similarly, if $q = 1$, the exchange operators $R^{\pm}_{VV}(u, \hat{\lambda})$ are gauge equivalent to the basic trigonometric solution without spectral parameter, with $q = e^{\pi i/m(k+g)}$; the gauge transformation involves an exact multiplicative 2-form expressing via classical Gamma-functions. This is obtained by sending q to 1 in the result of [Mo].*

Remark Note that the limit $q \to 1$ in this setting is rather subtle. Indeed, for $q \neq 1$ the function $J_{VW}(u, \lambda)$ has an infinite sequence of poles in the z-plane, which becomes denser as q approaches 1 and eventually degenerates into a branch cut; i.e. this single valued meromorphic function becomes multivalued in the limit.

3.11. The quantum Kazhdan-Lusztig functor

Let $\mathfrak{g} = \mathfrak{sl}_n$, $q \neq 1$. Example 3.6 allows one to construct a tensor functor from the category $\mathrm{Rep}_f(U_q(\hat{\mathfrak{g}}))$ of finite dimensional $U_q(\hat{\mathfrak{g}})$-modules, to the category $\mathrm{Rep}_f(R)$ of finite dimensional representations of the basic elliptic solution R of QDYBE with spectral parameter (i.e. to the category of finite dimensional representations of Felder's elliptic quantum group). Namely, let V be the vector representation, and for any finite dimensional representation W of $U_q(\hat{\mathfrak{g}})$, set $L_W = R_{VW}$. Then (W, L_W) is a representation of the dynamical R-matrix R_{VV}. This defines a functor $F : \mathrm{Rep}_f(U_q(\hat{\mathfrak{g}})) \to \mathrm{Rep}_f(R_{VV})$. Moreover, this functor is a tensor functor: the tensor structure $F(W) \otimes F(U) \to F(W \otimes U)$ is given by the fusion operator $J_{WU}(\hat{\lambda})$ (the axiom of a tensor structure follows from the dynamical twist equation for J). On the other hand, since R_{VV} and R are gauge equivalent by an exact form, their representation categories are equivalent, so one may assume that F lands in $\mathrm{Rep}_f(R)$.

If the scalars for $\mathrm{Rep}_f(U_q(\hat{\mathfrak{g}}))$ are taken to be the field of periodic functions of $\hat{\lambda}$ (in particular, k is regarded as a variable), then F is fully faithful; it can be regarded as a q-analogue of the Kazhdan-Lusztig functor (see [EM] and references therein). It generalizes to infinite dimensional representations, and allows one to

construct elliptic deformations of all evaluation representations of $U_q(\hat{\mathfrak{g}})$) (which was done for finite dimensional representations in [TV]). The versions of this functor without spectral parameter, from representations of $\mathfrak{g}$ ($U_q(\mathfrak{g})$) to representations of the basic rational (trigonometric) solution of QDYBE can be found in [EV3].

4. Traces of intertwining operators and Macdonald functions

In this section we discuss a connection between dynamical R-matrices and certain integrable systems and special functions (in particular, Macdonald functions).

4.1. Trace functions

Let V be a finite dimensional representation of $U_q(\mathfrak{g})$ ($q \neq 1$), such that $V[0] \neq 0$. Recall that for any $v \in V[0]$ and generic μ, one can define an intertwining operator Φ_μ^v such that $\langle \Phi_\mu^v \rangle = v$. Following [EV4] set $\Psi^v(\lambda, \mu) = \mathrm{Tr}|_{M_\mu}(\Phi_\mu^v q^{2\bar{\lambda}})$. This is an infinite series in the variables $q^{-(\lambda,\alpha_i)}$ (where α_i are the simple roots) whose coefficients are rational functions of $q^{(\mu,\alpha_i)}$ (times a common factor $q^{2(\lambda,\mu)}$). For generic μ this series converges near 0, and its matrix elements belong to $q^{2(\lambda,\mu)}(\mathbb{C}(q^{(\lambda,\alpha_i)}) \otimes \mathbb{C}(q^{(\mu,\alpha_i)}))$.

Let $\Psi_V(\lambda, \mu)$ be the $\mathrm{End}(V[0])$-valued function, such that $\Psi_V(\lambda, \mu)v = \Psi^v(\lambda, \mu)$. The function Ψ_V has remarkable properties and in a special case is closely related to Macdonald functions. To formulate the properties of Ψ_V, we will consider a renormalized version of this function. Namely, let $\delta_q(\lambda)$ be the Weyl denominator $\prod_{\alpha>0}(q^{(\lambda,\alpha)} - q^{-(\lambda,\alpha)})$. Let also $Q(\mu) = \sum S^{-1}(b_i)a_i$, where $\sum a_i \otimes b_i = J(\mu)$ is the universal fusion operator (this is an infinite expression, but it makes sense as a linear operator on finite dimensional representations; moreover it is of zero weight and invertible). Define the *trace function* $F_V(\lambda, \mu) = \delta_q(\lambda)\Psi_V(\lambda, -\mu - \rho)Q(-\mu - \rho)^{-1}$.

4.2. Commuting difference operators

For any finite dimensional $U_q(\mathfrak{g})$-module W, we define a difference operator $\mathcal{D}_W$ acting on functions on $\mathfrak{h}^*$ with values in $V[0]$. Namely, we set $(\mathcal{D}_W f)(\lambda) = \sum_{\nu \in \mathfrak{h}^*} \mathrm{Tr}|_W(R_{WV}(-\lambda - \rho))f(\lambda + \nu)$. These operators are dynamical analogs of transfer matrices, and were introduced in [FV2]. It can be shown that $\mathcal{D}_{W_1 \otimes W_2} = \mathcal{D}_{W_1}\mathcal{D}_{W_2}$; in particular, $\mathcal{D}_W$ commute with each other, and the algebra generated by them is the polynomial algebra in $\mathcal{D}_{W_i}$, where W_i are the fundamental representations of $U_q(\mathfrak{g})$.

4.3. Difference equations for the trace functions

It turns out that trace functions $F_V(\lambda, \mu)$, regarded as functions of λ, are common eigenfunctions of $\mathcal{D}_W$.

Theorem 4.1 [EV4] *One has*

$$\mathcal{D}_W^{(\lambda)} F_V(\lambda,\mu) = \chi_W(q^{-2\bar{\mu}}) F_V(\lambda,\mu), \tag{1}$$

where $\chi_W(x) = \mathrm{Tr}|_W(x)$ *is the character of* W.

In fact, it is easy to deduce from this theorem that if v_i is a basis of $V[0]$ then $F_V(\lambda,\mu)v_i$ is a basis of solutions of (1) in the power series space. Thus, trace functions allow us to integrate the quantum integrable system defined by the commuting operators $\mathcal{D}_{W_i}$.

Theorem 4.2 [EV4] *The function* F_V *is symmetric in* λ *and* μ *in the following sense:* $F_{V^*}(\mu,\lambda) = F_V(\lambda,\mu)^*$.

This symmetry property implies that F_V also satisfies "dual" difference equations with respect to μ: $\mathcal{D}_W^{(\mu)} F_V(\lambda,\mu)^* = \chi_W(q^{-2\bar{\lambda}}) F_V(\lambda,\mu)^*$.

4.4. Macdonald functions

An important special case, worked out in [EKi], is $\mathfrak{g} = \mathfrak{sl}_n$, and $V = L_{kn\omega_1}$, where ω_1 is the first fundamental weight, and k a nonnegative integer. In this case, $\dim V[0] = 1$, and thus trace functions can be regarded as scalar functions. Furthermore, it turns out ([FV2]) that the operators $\mathcal{D}_W$ can be conjugated (by a certain explicit product) to Macdonald's difference operators of type A, and thus the functions $F_V(\lambda,\mu)$ are Macdonald functions (up to multiplication by this product). One can also obtain Macdonald's polynomials by replacing Verma modules M_λ with irreducible finite dimensional modules L_λ; see [EV4] for details. In this case, Theorem 4.2 is the well known Macdonald's symmetry identity, and the "dual" difference equations are the recurrence relations for Macdonald's functions.

Remark 1 The dynamical transfer matrices $\mathcal{D}_W$ can be constructed not only for the trigonometric but also for the elliptc dynamical R-matrix; in the case $\mathfrak{g} = \mathfrak{sl}_n$, $V = L_{kn\omega_1}$ this yields the Ruijsenaars system, which is an elliptic deformation of the Macdonald system.

Remark 2 If $q = 1$, the difference equations of Theorem 4.1 become differential equations, which in the case $\mathfrak{g} = \mathfrak{sl}_n$, $V = L_{kn\omega_1}$ reduce to the trigonometric Calogero-Moser system. In this limit, the symmetry property is destroyed, but the "dual" difference equations remain valid, now with the exchange operator for $\mathfrak{g}$ rather than $U_q(\mathfrak{g})$. Thus, both for $q = 1$ and $q \neq 1$, common eigenfunctions satisfy additional difference equations with respect to eigenvalues – the so called bispectrality property.

Remark 3 Apart from trace Ψ^v of a single intertwining operator multiplied by $q^{2\bar{\lambda}}$, it is useful to consider the trace of a product of several such operators. After an appropriate renormalization, such multicomponent trace function (taking values in $\mathrm{End}((V_1 \otimes ... \otimes V_N)[0])$ satisfies multicomponent analogs of (1) and its dual version, as well as the symmetry. Furthermore, it satisfies an additional quantum Knizhnik-Zamolodchikov-Bernard equation, and its dual version (see [EV4]).

Remark 4 The theory of trace functions can be generalized to the case of affine Lie algebras, with V being a finite dimensional representation of $U_q(\hat{\mathfrak{g}})$. In this case, trace functions will be interesting transcendental functions. In the case $\mathfrak{g} = \mathfrak{sl}_n$, $V = L_{kn\omega_1}$, they are analogs of Macdonald functions for affine root systems. It is expected that for $\mathfrak{g} = \mathfrak{sl}_2$ they are the elliptic hypergeometric functions studied in [FV3]. This is known in the trigonometric limit ([EV4]) and for $q = 1$.

Remark 5 The theory of trace functions, both finite dimensional and affine, can be generalized to the case of any generalized Belavin-Drinfeld triple; see [ES1].

Remark 6 Trace functions $F_V(\lambda, \mu)$ are not Weyl group invariant. Rather, the diagonal action of the Weyl group multiplies them by certain operators, called the *dynamical Weyl group* operators (see [EV5]). These operators play an important role in the theory of dynamical R-matrices and trace functions, but are beyond the scope of this paper.

References

[ABRR] Arnaudon, D., Buffenoir, E., Ragoucy, E. and Roche, Ph., *Universal Solutions of quantum dynamical Yang-Baxter equations*, Lett. Math. Phys. **44** (1998), no. 3, 201–214.

[Dr] Drinfeld, V. G., Quantum groups, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986), 798–820, Amer. Math. Soc., Providence, RI, 1987.

[EFK] Etingof, P., Frenkel, I., Kirillov, A., Jr. *Lectures on representation theory and Knizhnik-Zamolodchikov equations*, AMS, (1998).

[EK] Etingof P., Kazhdan D., *Quantization of Lie bialgebras I*, Selecta Math., **2** (1996), 1–41.

[EKi] Etingof, P. I., Kirillov, A. A., Jr, *Macdonald's polynomials and representations of quantum groups*, Math. Res. Let., **1** (3) (1994), 279–296.

[EM] Etingof, P., Moura, A., On the quantum Kazhdan-Lusztig functor, math.QA/0203003, 2002.

[ES1] Etingof, P., Schiffmann, O., Twisted traces of quantum intertwiners and quantum dynamical R-matrices corresponding to generalized Belavin-Drinfeld triples, CMP 218 (2001), no. 3, 633–663.

[ES2] Etingof, P.; Schiffmann, O., Lectures on the dynamical Yang-baxter equations, math. QA/9908064.

[ESS] Etingof, P., Schedler, T., Schiffmann, O., *Explicit quantization of dynamical r-matrices*, J. Amer. Math. Soc., **13** (2000), 595–609.

[EV1] Etingof, P., Varchenko, A., *Geometry and classification of solutions of the classical dynamical Yang-Baxter equation*, Commun. Math. Phys, **192** (1998), 77–120 .

[EV2] Etingof, P., Varchenko, A., *Solutions of the quantum dynamical Yang-Baxter equation and dynamical quantum groups*, Commun. Math. Phys, **196** (1998), 591–640 .

[EV3] Etingof, P., Varchenko, A., *Exchange dynamical quantum groups*, CMP 205 (1999), no. 1, 19–52.

[EV4] Etingof, P., Varchenko, A., *Traces of intertwiners for quantum groups and difference equations, I*, Duke Math. J., 104 (2000), no. 3, 391–432.

[EV5] Etingof, P., Varchenko, A., *Dynamical Weyl groups and applications*, math.QA/0011001, to appear in Adv. Math.

[Fa] Faddeev, L., *On the exchange matrix of the WZNW model*, CMP, **132** (1990), 131–138.

[Fe] Felder, G., *Conformal field theory and integrable systems associated to elliptic curves*, Proceedings of the International Congress of Mathematicians, Zürich 1994, 1247–1255, Birkhäuser, 1994; *Elliptic quantum groups*, preprint hep-th/9412207, XIth International Congress of Mathematical Physics (Paris, 1994), 211–218, Internat. Press, Cambridge, MA, 1995.

[FR] Frenkel, I., Reshetikhin, N., *Quantum affine algebras and holonomic difference equations*, Commun. Math. Phys. **146** (1992), 1–60.

[FRT] Faddeev, L. D., Reshetikhin, N. Yu., Takhtajan, L. A., Quantization of Lie groups and Lie algebras. Algebraic analysis, Vol. I, 129–139, Academic Press, Boston, MA, 1988.

[FV1] Felder, G., Varchenko, A., *On representations of the elliptic quantum group* $E_{\tau,\eta}(sl_2)$, Commun. Math. Phys., **181** (1996), 746–762.

[FV2] Felder, G., Varchenko, A., *Elliptic quantum groups and Ruijsenaars models*, J. Statist. Phys., **89** (1997), no. 5–6, 963–980.

[FV3] Felder, G., Varchenko, A., The q-deformed Knizhnik-Zamolodchikov-Bernard heat equation, CMP 221 (2001), no. 3, 549–571.

[JKOS] Jimbo, M., Odake, S., Konno, H., Shiraishi, J., Quasi-Hopf twistors for elliptic quantum groups, Transform. Groups, 4 (1999), no. 4, 303–327.

[Lu] Lu, J. H., *Hopf algebroids and quantum groupoids*, Inter. J. Math., **7** (1) (1996), 47–70.

[Mo] Moura, A., Elliptic Dynamical R-Matrices from the Monodromy of the q-Knizhnik-Zamolodchikov Equations for the Standard Representation of Uq(sl(n+1)), math.RT/0112145.

[Sch] Schiffmann, O, *On classification of dynamical r-matrices*, MRL, **5** (1998), 13–30 .

[TV] Tarasov, V.; Varchenko, A. Small elliptic quantum group $e_{\tau,\gamma}(\mathfrak{sl}_N)$, Mosc. Math. J., 1 (2001), no. 2, 243–286, 303–304.

[Xu] Xu, P., Triangular dynamical r-matrices and quantization, Adv. Math., 166 (2002), no. 1, 1–49.

ICM 2002 · Vol. II · 571–582

Geometric Langlands Correspondence for GL_n

D. Gaitsgory*

Abstract

We will review the geometric Langlands theory (mainly for the group GL_n), whose development was initiated in the works of V. Drinfeld and G. Laumon.

Let X be a (smooth, complete) algebraic curve over a ground field k, and let E be an ℓ-adic n-dimensional irreducible local system on X.

The geometric Langlands conjecture (for GL_n) says that to E one can associate an automorphic sheaf $\mathcal{S}_E$, which is a perverse sheaf on the moduli stack $\mathrm{Bun}_n(X)$ classifying vector bundles of rank n on X.

We will explain the motivation for this conjecture in terms of the classical theory of automorphic forms, and the methods involved in the construction of $\mathcal{S}_E$.

2000 Mathematics Subject Classification: 14H60.

Introduction

0.1. Let X be a (smooth, complete) curve over a ground field k, and G a (split) reductive group. In the main body of the paper we will assume that $G = GL_n$, but now we would like to make a general overview of the theory, in which G can be arbitrary. Let Bun_G denote the moduli stack of G-bundles on X.

Our object of study is the category $\mathrm{D}(\mathrm{Bun}_G)$–the derived category of sheaves on Bun_G with constructible cohomology. By "sheaves" we will mean either $\overline{\mathbb{Q}}_\ell$-adic sheaves, which can be done over any k, or D-modules, when $\mathrm{char}(k) = 0$.

Finally, let $\check{G}$ be the Langlands dual of G (thought of as an algebraic group over $\overline{\mathbb{Q}}_\ell$ or k, depending on the sheaf-theoretic context).

0.2. It is believed that if σ is a $\check{G}$-local system on X (i.e., a homomorphism from the appropriate version of $\pi_1(X)$ to $\check{G}$), which is sufficiently generic, then to σ one can attach an *automotphic sheaf* $\mathcal{F}_\sigma \in \mathrm{D}(\mathrm{Bun}_G)$, which is a Hecke eigen-sheaf with respect to σ. (See [BD], Section 5 for the definition of the Hecke eigen-property for an arbitrary G, or Section 2 below for the GL_n case.)

*Department of Mathematics, The University of Chicago, 5734 University Ave., Chicago, IL 60637, USA. E-mail: gaitsgde@math.uchicago.edu

Unfortunately, the geometric Langlands conjecture, i.e., the conjecture predicting the existence of $\mathcal{F}_\sigma$, is not well formulated, because it is not known (at least to the author) what "sufficiently generic" is for an arbitrary group G.

The exception is the case when $G = GL_n$, and "sufficiently generic" is understood as "irreducible". In this situation, the geometric Langlands conjecture in the above form was formulated by G. Laumon in [La1], following the pioneering work of V. Drinfeld, [Dr], where the corresponding statement was proved for GL_2.

0.3. Although it is still not clear under what circumstances $\mathcal{F}_\sigma$ should exist, at least three different constructions have been suggested by A. Beilinson and V. Drinfeld, in addition to the original Drinfeld's construction (the latter is, however, specific to the case $G = GL_n$). To the best of my knowledge, there are no theorems that establish the equivalence between any two of the four constructions described below.

The first construction works in the D-module context (in particular, we have to assume that $\mathrm{char}(k) = 0$), under an additional assumption on σ: one assumes that σ is an *oper*, cf. [BD]. In this case, the corresponding $\mathcal{F}_\sigma$ is a D-module (and not just an object of the derived category). Moreover, $\mathcal{F}_\sigma$ is holonomic. (Unfortunately, it is not clear from the construction whether $\mathcal{F}_\sigma$ has regular singularities.)

The second construction, via the so-called *chiral Hecke algebra*, also takes place in the D-module context. Here σ can be arbitrary, but it is not clear under what assumptions on σ the object $\mathcal{F}_\sigma$ thus constructed is non-zero, or when it is a single D-module. It is not clear either whether the corresponding complex always has holonomic (or even finitely generated) cohomologies.

In the above two constructions, the fact that we work with D-modules is used very essentially, as the corresponding $\mathcal{F}_\sigma$ is constructed by generators and relations. The other two constructions are more geometric in the sense that $\mathcal{F}_\sigma$ is produced starting from σ, viewed as a sheaf on X, using the 6 functors.

The third construction, uses the "spectral projector", and makes sense over any k and for any σ. It is again not clear under what assumption on σ, the resulting $\mathcal{F}_\sigma$ is non-zero or when it lies in the bounded derived category.

0.4. Finally, the fourth approach which, as was mentioned above, works for GL_n only and goes back to the original work of Drinfeld, is the subject of the present paper. In this case it can actually be shown that for an irreducible representation σ, which can be thought of as an n-dimensional local system E on X, the corresponding automorphic sheaf $\mathcal{F}_\sigma$ (or $\mathcal{F}_E$) indeed exists and has all the desired properties.

Let us add a few words about the history of this approach. After [Dr], where the case of $n = 2$ was solved, the (conjectural) generalization of the construction of $\mathcal{F}_E$ was suggested by G. Laumon in [La1] and [La2]. Laumon's approach was further developed by E. Frenkel, D. Kazhdan, K. Vilonen and the author (cf. [FGKV] and [FGV1]). The present paper can be regarded as a summary of these works. Finally, a certain vanishing result that was missing in order to complete the proof of the conjecture has been established in [Ga].

Let us now briefly explain how the paper is organized. In Section 1 we review the classical (i.e., function-theoretic) Langlands correspondence for GL_n. In Section

2 we define Hecke eigen-sheaves and state the main theorem about the existence of the Hecke eigen-sheaf $\mathcal{F}_E$ corresponding to an irreducible local system E. In Section 3 we describe the construction of $\mathcal{F}_E$ via a geometric analog of the Whittaker model. In Section 4 we explain how the main theorem about the existence of $\mathcal{F}_E$ follows from a certain vanishing result. Finally, in Section 5 we indicate the steps involved in the proof of the vanishing theorem.

1. The classical theory

In this section we will review the formulation of the classical Langlands conjecture for function fields, and the technique of construction of automorphic forms via Whittaker models.

1.1. Let $\mathcal{K}$ be the global field corresponding to a (smooth, complete) curve X over a finite field $\mathbb{F}_q$. We will denote by $\mathbb{A}$ (resp., $\mathbb{O}$) the corresponding ring of adeles (resp., the subring of integral adeles).

Consider the quotient $GL_n(\mathcal{K})\backslash GL_n(\mathbb{A})$. The space $\mathrm{Funct}(GL_n(\mathcal{K})\backslash GL_n(\mathbb{A}))$ of (smooth) functions (with values in an arbitrary algebraically closed field of char. 0, which we will take to be $\overline{\mathbb{Q}}_\ell$) is naturally a representation of the group $GL_n(\mathbb{A})$.

Consider the subspace of functions that are invariant with respect to $GL_n(\mathbb{O})$, i.e. the space of functions on the double quotient $GL_n(\mathcal{K})\backslash GL_n(\mathbb{A})/GL_n(\mathbb{O})$. This is a module over the Hecke algebra $GL_n(\mathbb{A})$ with respect to $GL_n(\mathbb{O})$. By definition, this Hecke algebra, denoted $\mathrm{H}(GL_n(\mathbb{A}), GL_n(\mathbb{O}))$, consists of compactly supported $GL_n(\mathbb{O})$-biinvariant functions on $GL_n(\mathbb{A})$. The action of $h \in \mathrm{H}(GL_n(\mathbb{A}), GL_n(\mathbb{O}))$ on $f \in \mathrm{Funct}(GL_n(\mathcal{K})\backslash GL_n(\mathbb{A})/GL_n(\mathbb{O}))$ is given by the formula

$$h \cdot f(x) = \int\limits_{y \in GL_n(\mathbb{A})} f(x \cdot y^{-1}) \cdot h(y). \tag{1.1}$$

By the definition of the adele group $GL_n(\mathbb{A})$, the Hecke algebra $\mathrm{H}(GL_n(\mathbb{A}), GL_n(\mathbb{O}))$ is the (restricted) tensor product of local Hecke algebras $\otimes'_x \mathrm{H}(GL_n(\mathcal{K}_x), GL_n(\mathcal{O}_x))$, where the index x runs over the set of all places of X, and $\mathcal{K}_x$ (resp., $\mathcal{O}_x$) is the corresponding local field (resp., local ring).

The structure of each local Hecke algebra $\mathrm{H}(GL_n(\mathcal{K}_x), GL_n(\mathcal{O}_x))$ is well understood. It is commutative an is freely generated by the elements $T_x^1, T_x^2, ..., T_x^n, (T_x^n)^{-1}$, where each is T_x^i the characteristic function of the $GL_n(\mathcal{O}_x)$-double coset in $GL_n(\mathcal{K}_x)$ corresponding to the diagonal matrix $(\underbrace{\varpi_x, ..., \varpi_x}_{i \text{ times}}, 1, ..., 1)$, where ϖ_x is the uniformizer at x. (It is convenient to normalize T_x^i so that its value is not 1 but rather $(-q^{-1/2})^{i(n-i)}$.)

1.2. Let σ be an n-dimensional representation of the unramified quotient of the Galois group of $\mathcal{K}$. To σ we can attach a canonical character χ_σ of $\mathrm{H}(GL_n(\mathbb{A}), GL_n(\mathbb{O}))$. Namely, the value of χ_σ on T_x^i, $\forall x, i$ is $\mathrm{Tr}(\Lambda^i \mathrm{Frob}_x)$.

Here is the formulation of the Galois $\Rightarrow$ Automorphic part of the Langlands correspondence, established by Lafforgue in [Laf]:

1.3. Theorem. *To every irreducible σ as above there corresponds a (non-zero) function $f_\sigma \in \mathrm{Funct}(GL_n(\mathcal{K})\backslash GL_n(\mathbb{A})/GL_n(\mathbb{O}))$, such that*

$$h \cdot f_\sigma = \chi_\sigma(h) \cdot f_\sigma, \ \forall h \in \mathrm{H}(GL_n(\mathbb{A}), GL_n(\mathbb{O})).$$

Moreover, such f_σ is unique up to a scalar, and is cuspidal (see below).

1.4. We will now sketch the construction of f_σ using the method of Piatetski-Shapiro and Shalika, cf. [PS],[Sha].

First, let us recall the notion of cuspidality of a function on $GL_n(\mathbb{A})$. Let $f \in \mathrm{Funct}(GL_n(\mathbb{A}))$ be left-invariant with respect to a subgroup $\Gamma(\mathcal{K})$, where $\Gamma \subset GL_n$ is a subgroup that contains the unipotent radical N of the standard Borel subgroup. In particular, Γ contains also the unipotent radical $N(Q)$ of any standard parabolic $Q \subset GL_n$. The function f is called cuspidal if for every such parabolic

$$\int\limits_{y \in N(Q)(\mathcal{K})\backslash N(Q)(\mathbb{A})} f(y \cdot x) = 0, \ \forall x \in GL_n(\mathbb{A}). \tag{1.2}$$

For $k = 1, ..., n$ let $P_k \subset GL_n$ be the group of matrices, for which the a_{ij} entry is δ_{ij} if $i > n - k, i \geq j$. Note that $P := P_1$ is the so-called mirabolic subgroup, i.e. the subgroup of matrices whose last row is $(0, ..., 0, 1)$, and for any k, $P_k \supset N$.

Let $N_k \subset GL_n$ be the subgroup of (strictly) upper-triangular matrices, in which only the last $k-1$ columns may be non-zero. We shall fix a non-trivial character $\psi : \mathcal{K}\backslash\mathbb{A} \to \overline{\mathbb{Q}}_\ell^*$. It gives rise to a character $\Psi_k : N_k(\mathbb{A}) \to \overline{\mathbb{Q}}_\ell^*$, but taking the sum of values of ψ on the supdiagonal matrix entries.

We define the space W_k to consist of all functions

$$f \in \mathrm{Funct}(P_i(\mathcal{K})\backslash GL_n(\mathbb{A})/GL_n(\mathbb{O})),$$

that satisfy $f(n \cdot x) = \Psi_k(n) \cdot f(x)$, $\forall n \in N_k$. We define $\mathrm{W}_{cusp\, k}$ as the subspace of W_k that corresponds to cuspidal functions. It is easy to see that for $k = n$, $\mathrm{W}_{cusp\, n} = \mathrm{W}_n$.

1.5. Proposition. *There are isomorphisms $\mathrm{W}_{cusp\, k} \simeq \mathrm{W}_{cusp\, k+1}$ for $k = 1, ..., n-1$, which respect the $\mathrm{H}(GL_n(\mathbb{A}), GL_n(\mathbb{O}))$-action.*

The isomorphisms in the proposition are given by taking Fourier transforms along the (compact abelian) group $(N_{k+1}/N_k)(\mathcal{K})\backslash(N_{k+1}/N_k)(\mathbb{A})$.

1.6. The construction of the sought-for automorphic function f_σ corresponding to a Galois representation σ goes as follows. One first constructs (using an explicit formula due to Shintani, [Shi], and Casselman–Shalika, [CS]) the corresponding Whittaker function, i.e., a function $f_\sigma^W \in \mathrm{W}_n$, on which the Hecke algebra $\mathrm{H}(GL_n(\mathbb{A}), GL_n(\mathbb{O}))$ acts via the character χ_σ.

Then we use the sequence if isomorphisms of Proposition 1.5. for $k = n-1, ..., 2, 1$ and obtain a cuspidal function f'_σ on $P(\mathcal{K})\backslash GL_n(\mathbb{A})$. The crux of the matter is to show that f'_σ comes in fact from a $GL_n(\mathcal{K})$-invariant function, which would be the function f_σ we were looking for. In his work, Lafforgue deduces the existence of such f_σ from the functional equation satisfied by the Rankin-Selberg L-functions corresponding to σ and automorphic forms on $GL_{n'}$ with $n' < n$.

2. The geometric setting

In this section we will give a formulation of the geometric Langlands correspondence.

We will assume that the ground field k is of positive characteristic and is algebraically closed. (The case of D-modules over a ground field of char. 0 is completely analogous.) We will denote by Bun_n be the moduli stack of rank n vector bundles on our curve X; $\mathrm{D}(\mathrm{Bun}_n)$ will denote the corresponding derived category of sheaves on Bun_n.

2.1. First, we will introduce the Hecke functors that act on the category $\mathrm{D}(\mathrm{Bun}_n)$. For a point $x \in X$ consider the stack $\mathcal{H}_x$ classifying the data of $(\mathcal{M}, \mathcal{M}', \beta)$, where $\mathcal{M}, \mathcal{M}'$ are vector bundles on X, and β is an embedding of coherent sheaves $\mathcal{M}' \hookrightarrow \mathcal{M}$, such that the quotient $\mathcal{M}/\mathcal{M}'$ is isomorphic to the residue field k_x, viewed as a coherent sheaf on X.

We have two projections: $\mathrm{Bun}_n \overset{\overleftarrow{h}}{\longleftarrow} \mathcal{H}_x \overset{\overrightarrow{h}}{\longrightarrow} \mathrm{Bun}_n$, that send the data of $(\mathcal{M}, \mathcal{M}', \beta)$ to $\mathcal{M}$ and $\mathcal{M}'$, respectively.

The (local) Hecke functor $\mathrm{H}_x : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(\mathrm{Bun}_n)$ is defined by the formula

$$\mathcal{F} \mapsto \overrightarrow{h}_!(\overleftarrow{h}^*(\mathcal{F}))[n-1]. \tag{2.1}$$

The above definition can be generalized by letting the point x be a parameter. We introduce the stack $\mathcal{H}$ fibered over X, whose fiber over each $x \in X$ is $\mathcal{H}_x$. We will denote the map $\mathcal{H} \to X$ by s. Thus, we have a diagram:

$$\mathrm{Bun}_n \overset{\overleftarrow{h}}{\longleftarrow} \mathcal{H}_x \overset{s \times \overrightarrow{h}}{\longrightarrow} X \times \mathrm{Bun}_n,$$

and we introduce the Hecke functor $\mathrm{H} : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(X \times \mathrm{Bun}_n)$ by $\mathcal{F} \mapsto (s \times \overrightarrow{h})_!(\overleftarrow{h}^*(\mathcal{F}))[n]$.

2.2. For an integer d, let us denote by $\mathrm{H}^{\boxtimes d}$ the corresponding iteration of H, i.e.

$$(\mathrm{id}_{X^{d-1}} \boxtimes \mathrm{H}) \circ ... (\mathrm{id}_X \boxtimes \mathrm{H}) \circ \mathrm{H} : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(X^d \times \mathrm{Bun}_n). \tag{2.2}$$

This functor has the following basic property:

2.3. Proposition. *The functor* $\mathrm{H}^{\boxtimes d} : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(X^d \times \mathrm{Bun}_n)$ *lifts in a natural way to a functor into the equivariant derived category* $\mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}^{\Sigma_d}(X^d \times \mathrm{Bun}_n)$, *where* Σ_d *is the symmetric group that acts naturally on* X^d.

2.4. We are now ready to define what a Hecke eigen-sheaf is. Let E be an n-dimensional local system on X.

An object $\mathcal{F}_E \in \mathrm{D}(\mathrm{Bun}_n)$ is called a Hecke eigen-sheaf with respect to E, if we are given an isomorphism

$$\mathrm{H}(\mathcal{F}_E) \simeq E[1] \boxtimes \mathcal{F}_E,$$

such that the induced map

$$\mathrm{H}^{\boxtimes d}(\mathcal{F}_E) \to E^{\boxtimes d}[d] \boxtimes \mathcal{F}_E$$

respects the Σ_d-equivariant structure for any d. [1]

Few remarks are in order:

Remark 1. In addition to the "basic" Hecke functor H, one can introduce the functors $\mathrm{H}^i : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(X \times \mathrm{Bun}_n)$ for $i = 1, ..., n$ (classically, they correspond to the generators T_x^i of the Hecke algebra of $\mathrm{H}(GL_n(\mathcal{K}_x), GL_n(\mathcal{O}_x))$). We have $\mathrm{H} \simeq \mathrm{H}^1$, and H^n is the pull-back functor with respect to the multiplication map $(x, \mathcal{M}) \to \mathcal{M}(x)$. One can show (cf. [FGV1], Sect. 2) that if $\mathcal{F}_E$ is a Hecke eigen-sheaf with respect to E, then $\mathrm{H}^i(\mathcal{F}_E) \simeq \Lambda^i(E) \boxtimes \mathcal{F}_E$.

Remark 2. Note, that formally in the definition of eigen-sheaves, we did not use the fact that the local system E was n-dimensional. However, one can show (using Remark 1 above) that if $\mathcal{F}_E \in \mathrm{D}(\mathrm{Bun}_n)$ is a Hecke eigen-sheaf with respect to E and the rank of E is different from n, then $\mathcal{F}_E = 0$.

2.5. The following is the statement of the (unramified) geometric Langlands correspondence for GL_n, conjectured by G. Laumon in [La1] and proved in [FGV1] and [Ga]:

2.6. Theorem. *Suppose that the local system E is irreducible. Then there exists a Hecke eigen-sheaf $\mathcal{F}_E \in \mathrm{D}(\mathrm{Bun}_n)$ with respect to E, which is, moreover, an irreducible perverse sheaf over every connected component of* Bun_n, *and cuspidal.* [2]

Remark 3. Of course, if $\mathcal{F}_E$ is a Hecke eigen-sheaf, then so is $\mathcal{F}_E \otimes K$, where K is any complex of vector spaces. An additional conjecture, which has not been fully established yet, is that any Hecke eigen-sheaf with respect to E has this form, where $\mathcal{F}_E$ is *the* eigen-sheaf constructed in [FGV1].

3. Geometric Whittaker models

From now on our goal will be to sketch the steps involved in the construction of $\mathcal{F}_E$.

3.1. Let Bun_n' be the stack classifying the data of $(\mathcal{M}, \kappa)$, where $\mathcal{M}$ is a rank n vector bundle, and κ is a non-zero map $\Omega^{n-1} \to \mathcal{M}$. [3] Let π denote the natural projection $\mathrm{Bun}_n' \to \mathrm{Bun}_n$.

We will construct an object $\mathcal{F}_E' \in \mathrm{D}(\mathrm{Bun}_n')$, and then show that it descends to the sought-for perverse sheaf $\mathcal{F}_E$ on Bun_n.

For us, the category $\mathrm{D}(\mathrm{Bun}_n')$ is the geometric analog of the space of functions on the quotient $P(K)\backslash GL_n(\mathbb{A})/GL_n(\mathbb{O})$, and the construction of $\mathcal{F}_E'$ will be an analog of the construction of the corresponding automorphic function f_σ' (cf. Section 1) from its Whittaker model, i.e. f_σ^W. Thus, we must find stacks and certain subcategories of sheaves on them, that will be analogs of the "Whittaker" spaces W_k of Section 1.

[1] The latter condition makes sense, since the forgetful functor from the equivariant derived category with respect to a finite group to the usual derived category is faithful, because the coefficients of our sheaves are in char. 0.

[2] See [FGV1], Section 9 for the notion of cuspidality.

[3] Here and in the sequel, Ω^k denotes the k-th tensor power of the canonical line bundle Ω.

3.2. For an integer k, $1 \leq k \leq n$, let $\overline{\mathcal{Q}}_{n,k}$ be the stack classifying the data of $(\mathcal{M}, \kappa_1, ..., \kappa_k)$, where $\mathcal{M}$ is a rank n vector bundle, κ_i's are non-zero maps $\Omega^{n-1+...+n-i} \to \Lambda^k(\mathcal{M})$, satisfying the Plücker relations. The latter means that at the generic point of X, the collection of maps $\kappa_1, ..., \kappa_k$ defines a flag of subbundles

$$0 = \mathcal{M}_0 \subset \mathcal{M}_1 \subset ... \subset \mathcal{M}_k \subset \mathcal{M},$$

with identifications $\mathcal{M}_i/M_{i-1} \simeq \Omega^{n-i}$ (cf. [BG], Sect. 1.3 for details). This flag would be defined over the entire X if the maps κ_i were bundle maps, and not just injective maps of sheaves.

In addition, let $\overline{\mathcal{Q}}_{n,k,ex}$ be a slightly larger stack, where the last map κ_k is allowed to vanish. Let us denote by j_k the open embedding $j_k : \overline{\mathcal{Q}}_{n,k} \to \overline{\mathcal{Q}}_{n,k,ex}$. In addition, we have the natural maps $\pi_{k+1,k} : \overline{\mathcal{Q}}_{n,k+1} \to \overline{\mathcal{Q}}_{n,k}$ (resp., $\pi_{k+1,k,ex} : \overline{\mathcal{Q}}_{n,k+1,ex} \to \overline{\mathcal{Q}}_{n,k}$), which "forget" the data of κ_{k+1}.

3.3. Next, we claim that for each $k = 1, ..., n$ there exist a certain full subcategory $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k}) \subset \mathrm{D}(\overline{\mathcal{Q}}_{n,k})$, which is a geometric analog of the space W_k.

The definition of $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k})$ involves an action on $\overline{\mathcal{Q}}_{n,k}$ of a certain natural groupoid, which comes from the same unipotent group N_k as the one used in the definition of Whittaker functions on $P_k(\mathcal{K})\backslash GL_n(\mathbb{A})/GL_n(\mathbb{O})$. We refer the reader to [FGV] or [Ga], Sections 4 and 5 for a detailed discussion. Note that for $k = 1$, $\overline{\mathcal{Q}}_{n,1}$ is nothing but the stack Bun'_n introduced above, and in this case $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,1}) = \mathrm{D}(\overline{\mathcal{Q}}_{n,1}) = \mathrm{D}(\mathrm{Bun}'_n)$.

The relation between the categories $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k})$ for different k's is given by the next proposition. Before stating it, we should mention that in addition to $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k})$, we have the corresponding full subcategory $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k,ex}) \in \mathrm{D}(\overline{\mathcal{Q}}_{n,k,ex})$ and the functors $j_{k!}, j_{k*}, j_k^*$ map the categories $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,i,ex})$ and $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k})$ to one-another.

3.4. Proposition. *The direct image functor $\pi_{k+1,k,ex!} : \mathrm{D}(\overline{\mathcal{Q}}_{n,k+1,ex}) \to \mathrm{D}(\overline{\mathcal{Q}}_{n,k})$ maps the subcategory $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k+1,ex})$ to $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k})$ and induces an equivalence between the latter categories. Moreover, $\pi_{k+1,k,ex!}$ restricted to $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,i+1,ex})$ coincides with $\pi_{k+1,k,ex*}$; hence it preserves the subcategory of perverse sheaves and commutes with the Verdier duality.*

We introduce the Whittaker functor $W_{k,k+1} : \mathrm{D}^W(\overline{\mathcal{Q}}_{n,k}) \to \mathrm{D}^W(\overline{\mathcal{Q}}_{n,k+1})$ as a composition of the quasi-inverse of $\pi_{k+1,k,ex!}$ followed by the restriction j_k^*. The left (resp., right) adjoint of $W_{k,k+1}$, is nothing but $\pi_{k+1,k!}$ (resp., $\pi_{k+1,k*}$).

3.5. The object $\mathcal{F}'_E \in \mathrm{D}(\mathrm{Bun}'_n)$, that we promised to construct in this section is obtained as follows: $\mathcal{F}'_E$ is by definition

$$\pi_{2,1!} \circ \pi_{3,2!} \circ ... \pi_{n,n-1!}(\mathcal{F}^W_E), \tag{3.1}$$

where $\mathcal{F}^W_E$ is a certain canonical perverse sheaf in $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,n})$, that we will presently describe.

Remark 4. The above definition of $\mathcal{F}'_E$ can be rewritten as $\mathcal{F}'_E \simeq \mathsf{f}_!(\mathcal{F}^W_E)$, where f is the natural map $\overline{\mathcal{Q}}_{n,n} \to \mathrm{Bun}'_n$. However, we had to break $\mathsf{f}_!$ into the above

elementary steps (i.e. $\pi_{k+1,k!}$) in order to control perversity and irreducibility, which is crucially important, as we will see in the next section.

Let $\mathcal{Q}_{n,n} \subset \overline{\mathcal{Q}}_{n,n}$ be the open substack corresponding to the condition that the maps $\kappa_1, ..., \kappa_{n-1}$ are bundle maps. Let $\overset{\circ}{j}$ denote the open embedding of $\mathcal{Q}_{n,n}$ into $\overline{\mathcal{Q}}_{n,n}$. There are two natural maps defined on $\mathcal{Q}_{n,n}$.

One is the map, denoted $\tau : \mathcal{Q}_{n,n} \to \underset{d \geq 0}{\cup} X^{(d)}$, where $X^{(d)}$ denoted the d-th symmetric power of X. This map sends the data of $(\mathcal{M}, \kappa_1, ..., \kappa_n)$ to the divisor of zeroes of the map κ_n, which, we recall, is a map between line bundles.

The second map is denoted by $\mathrm{ev} : \mathcal{Q}_{n,n} \to \mathbb{G}_a$, and it is defined as follows. By the definition of $\mathcal{Q}_{n,n}$, a point of this stack defines a complete flag of vector bundles $0 = \mathcal{M}_0 \subset \mathcal{M}_1 \subset ... \subset \mathcal{M}_n = \mathcal{M}$, and identifications $\mathcal{M}_i/\mathcal{M}_{i-1} \simeq \Omega^{n-i}$ for $i < n$, and a map $\mathcal{O} \to \mathcal{M}_n/\mathcal{M}_{n-1}$ for $i = n$. Therefore, we have $n-1$ short exact sequences $0 \to \Omega^{j+1} \to \mathcal{M}^j \to \Omega^j \to 0$, where $\mathcal{M}^j$ is the corresponding rank 2 subquotient of $\mathcal{M}$, and each such extension defines a class in $H^1(X, \Omega) \simeq \mathbb{G}_a$. The value of ev on the above point of $\mathcal{Q}_{n,n}$ is the sum of the above extension $n-1$ classes in $\mathbb{G}_a$.

We define the perverse sheaf $\overset{\circ}{\mathcal{F}}{}^W_E$ on $\mathcal{Q}_{n,n}$ as a tensor product of $\tau^*(\underset{d}{\cup} E^{(d)})$ (here $E^{(d)}$ denotes the symmetric power of E which lives on $X^{(d)}$) and $\mathrm{ev}^*(\text{A-Sch})$, where A-Sch is the Artin-Schreier sheaf on $\mathbb{G}_a$. We apply an appropriate cohomological shift to the above tensor product to make it a perverse sheaf.

Finally, we define $\mathcal{F}^W_E \in \mathrm{D}(\overline{\mathcal{Q}}_{n,i})$ a s Goresky-MacPherson extension of $\overset{\circ}{\mathcal{F}}{}^W_E$, i.e. $\mathcal{F}^W_E := \overset{\circ}{j}_{!*}(\overset{\circ}{\mathcal{F}}{}^W_E)$. The fact that $\mathcal{F}^W_E$ belongs to the Whittaker category $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,n})$ follows from the construction.

3.6. In the next section we will explain how the irreducibility assumption on E implies that the complex $\mathcal{F}'_E$ on Bun'_n, constructed in this way, descends to Bun_n. Here we will comment on the action of the Hecke functors.

In [Ga], Section 6, it was shown that one can lift the Hecke functors H_x and H from Bun_n on the stacks $\overline{\mathcal{Q}}_{n,k}$.

More precisely, for every k we have the functors $\mathrm{H}_x^{\overline{\mathcal{Q}}_{n,k}} : \mathrm{D}(\overline{\mathcal{Q}}_{n,k}) \to \mathrm{D}(\overline{\mathcal{Q}}_{n,k})$, which preserve the subcategory $\mathrm{D}^W(\overline{\mathcal{Q}}_{n,k})$. Moreover, for $k = 2, ..., n$, the functors $W_{k-1,k}, \pi_{k,k-1!}, \pi_{k,k-1*}$ intertwine in the natural sense the functors $\mathrm{H}_x^{\overline{\mathcal{Q}}_{n,k}}$ and $\mathrm{H}_x^{\overline{\mathcal{Q}}_{n,k-1}}$. For $n = 1$, the pull-back functor (with an appropriate cohomological shift) π^* intertwines $\mathrm{H}_x^{\overline{\mathcal{Q}}_{n,1}}$ acting on $\mathrm{D}(\mathrm{Bun}'_n)$ and H_x acting on $\mathrm{D}(\mathrm{Bun}_n)$.

Analogusly, we have the global Hecke functors $\mathrm{H}^{\overline{\mathcal{Q}}_{n,k}} : \mathrm{D}(\overline{\mathcal{Q}}_{n,k}) \to \mathrm{D}(X \times \overline{\mathcal{Q}}_{n,k})$, where the assertions similar to the above ones hold.

The basic fact about the perverse sheaf $\mathcal{F}^W_E$ is its eigen-property, cf. [FGV1], Appendix:

3.7. Proposition. *There is a canonical isomorphism* $\mathrm{H}^{\overline{\mathcal{Q}}_{n,n}}(\mathcal{F}^W_E) \simeq E[1] \boxtimes \mathcal{F}^W_E$, *compatible with iterations (cf. Section 2).*

From this proposition it follows that $\mathcal{F}'_E$ satisfies the Hecke eigen-property with respect to $\mathrm{H}^{\overline{\mathcal{Q}}_{n,1}}$. This implies that the complex $\mathcal{F}_E \in \mathrm{D}(\mathrm{Bun}_n)$, to which $\mathcal{F}'_E$ descends, has the required Hecke eigen-property with respect to H.

4. Irreducibility and descent

The main step in the proof of Theorem 2.6. is the following theorem:

Recall that $\mathcal{F}'_E$ was obtained from the perverse sheaf $\mathcal{F}^W_E$ by applying to it the functor $\pi_{2,1!} \circ \pi_{3,2!} \circ ... \pi_{n,n-1!}$. Set $\mathcal{F}^n_E := \mathcal{F}^W_E$ and $\mathcal{F}^k_E := \pi_{k+1,k!} \circ ... \pi_{n,n-1!}(\mathcal{F}^W_E)$.

Let also $\mathrm{Bun}^{st}_n \subset \mathrm{Bun}_n$ be the open substack corresponding to vector bundles $\mathcal{M}$, for which $\mathrm{Hom}(\mathcal{L}^0, \mathcal{M}) = 0$ for some fixed line bundle $\mathcal{L}$. For every k we then obtain an open substack $\overline{\mathcal{Q}}^{st}_{n,k} := \mathrm{Bun}^{st}_n \underset{\mathrm{Bun}_n}{\times} \overline{\mathcal{Q}}_{n,k} \subset \overline{\mathcal{Q}}_{n,k}$.

4.1. Theorem. *Assume that E is irreducible. Then for $k = n, ..., 2$ the natural map $j_{k!}(\mathcal{F}^k_E) \to j_{i*}(\mathcal{F}^k_E)$ is an isomorphism over $\overline{\mathcal{Q}}^{st}_{n,k}$.*

From this theorem, we obtain that (over $\overline{\mathcal{Q}}^{st}_{n,k}$) $\mathcal{F}^k_E$ is a perverse sheaf, which is irreducible on every connected component. Indeed, this is true for $k = n$ by construction, and by induction we can assume that this is true also for $k' > k$. But then we obtain that $\mathcal{F}^k_E \simeq \pi_{k+1,k,ex!}(j_{k!*}(\mathcal{F}^{k+1}_E))$, and we know from Proposition 3.4. that the functor $\pi_{k+1,k,ex!}$ preserves perversity and irreducibility on the Whittaker subcategory. In particular, we obtain that $\mathcal{F}'_E$ is a perverse sheaf, irreducible on every connected component of $\mathrm{Bun}^{st\,\prime}_n$.

4.2. Theorem 4.1. has been established in [FGV1], in a slightly different form. The main ingredient in the proof of Theorem 4.1. is the following vanishing result:

For any local system E', and positive integers n' and d, let us consider the following functor $\mathrm{Av}^d_{n',E'} : \mathrm{D}(\mathrm{Bun}_{n'}) \to \mathrm{D}(\mathrm{Bun}_{n'})$:

Let $\mathrm{Mod}^d_{n'}$ be the stack of "upper modifications of length d", i.e. $\mathrm{Mod}^d_{n'}$ classifies the data of triples $(\mathcal{M}, \mathcal{M}', \beta)$, where $\mathcal{M}$ and $\mathcal{M}'$ are rank-n' vector bundles on X, and β is an embedding $\mathcal{M}' \hookrightarrow \mathcal{M}$ as coherent sheaves. Let $\overleftarrow{h}$ and $\overrightarrow{h}$ be the two projections from $\mathrm{Mod}^d_{n'}$ to $\mathrm{Bun}_{n'}$. In addition, there is a natural map $\mathfrak{s}$ from $\mathrm{Mod}^d_{n'}$ to the stack Coh^d_0 classifying torsion sheaves of length d on X. The map $\mathfrak{s}$ sends $(\mathcal{M}, \mathcal{M}', \beta)$ to the quotient $\mathcal{M}/\mathcal{M}'$.

Following Laumon [La1] (see also [FGV1] and [Ga]), starting from E' one defines Laumon's perverse sheaf $\mathcal{L}^d_{E'}$ on Coh^d_0. Set

$$\mathrm{Av}^d_{n',E'}(\mathcal{F}) := \overrightarrow{h}_!(\mathfrak{s}^*(\mathcal{L}^d_{E'}) \otimes \overleftarrow{h}^*(\mathcal{F})). \tag{4.1}$$

4.3. Theorem. *Suppose that E' is irreducible and that $\mathrm{rk}(E') > n'$. Suppose also that $d > (2g-2) \cdot \mathrm{rk}(E') \cdot n'$, where g is the genus of X. Then the functor $\mathrm{Av}^d_{n',E'} : \mathrm{D}(\mathrm{Bun}_{n'}) \to \mathrm{D}(\mathrm{Bun}_{n'})$ vanishes identically.*

The idea of the proof of Theorem 4.1. is that the cone of the map $j_{k!}(\mathcal{F}^k_E) \to j_{k*}(\mathcal{F}^k_E)$ can be expressed in terms of the functors $\mathrm{Av}^d_{k,E}$ for various d applied to various sheaves on Bun_{n-k+1}.

4.4. The next step is to show that perversity and irreducibility of $\mathcal{F}'_E$ on $\mathrm{Bun}_n^{st\,\prime}$ implies that $\mathcal{F}'_E$ descends to a perverse sheaf on Bun_n, cf. [FGV1], Section 5.

First, one shows that $\mathcal{F}'_E|_{\mathrm{Bun}_n^{st\,\prime}}$ descends to a perverse sheaf on Bun_n^{st}. This is done using a trick with Euler characteristics. It turns out, that since we already know that $\mathcal{F}'_E|_{\mathrm{Bun}_n^{st\,\prime}}$ is perverse and irreducible, it is sufficient to show that the Euler-Poincaré characteristics of the stalks of $\mathcal{F}'_E$ are constant along the fibers of the map $\pi : \mathrm{Bun}'_n \to \mathrm{Bun}_n$. Secondly, one observes that the above Euler-Poincaré characteristics are actually independent of the local system E (and depend only on its rank).

Thus, it is sufficient to find just one local system of a given rank n, for which the above constancy of Euler-Poincaré characteristics takes place, and one easily finds one like that.

Finally, using the Hecke eigen-property of $\mathcal{F}'_E$, one shows that there exists a perverse sheaf $\mathcal{F}_E$ defined on the entire Bun_n, whose (cohomologically normalized) pull-back to Bun'_n is isomorphic to $\mathcal{F}'_E$. The fact that the resulting sheaf $\mathcal{F}_E$ is cuspidal follows from the construction.

5. The vanishing result

In this section we will indicate the main ideas involved in the proof of Theorem 4.3., following [Ga]. We will change the notation slightly, and replace n' by n and E' by E.

5.1. First, let us rewrite the definition of the functor $\mathrm{Av}^d_{n,E}$. Consider first the functor $\mathrm{Av}_{n,E} : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(\mathrm{Bun}_n)$ given by

$$\mathcal{F} \mapsto p_!(q^*(E') \otimes H(\mathcal{F})),$$

where p, q are the two projections from $X \times \mathrm{Bun}_n$ to Bun_n and X, respectively.

From Proposition 2.3. it follows that the d-fold iteration $\mathrm{ItAv}^d_{n,E} := \mathrm{Av}_{n,E} \circ ... \circ \mathrm{Av}_{n,E'} : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(\mathrm{Bun}_n)$ maps to the equivariant derived category $\mathrm{D}^{\Sigma_d}(\mathrm{Bun}_n)$, where the Σ_d-action on the base Bun_n is, of course, trivial.

Moreover, it follows from the definition of Laumon's sheaf $\mathcal{L}^d_E$, that there is a functorial isomorphism

$$\mathrm{Av}^d_{n,E}(\mathcal{F}) \simeq (\mathrm{ItAv}^d_{n,E}(\mathcal{F}))^{\Sigma_d}, \tag{5.1}$$

where the superscript Σ_d designates the functor of Σ_d-invariants.

5.2. The first step in the proof of Theorem 4.3. is the observation that instead of proving that the functor $\mathrm{Av}_{n,E}$ *vanishes*, it is in fact enough to show that it is exact in the sense of the perverse t-structure.

The fact that the seemingly weaker exactness assertion is equivalent to vanishing is proved using the Euler characteristics trick, similar to what we did in the previous section.

Since, as we have seen above, the functor $\mathrm{Av}^d_{n,E}$ can be expressed through more elementary functors $\mathrm{Av}_{n,E}$, we will analyze the exactness properties of the

latter. Unfortunately, it is not true that the "elementary" functor $\mathrm{Av}_{n,E}$ is exact. However, it will be exact when regarded as a functor acting on a certain quotient category.

5.3. We introduce the category $\widetilde{\mathrm{D}}(\mathrm{Bun}_n)$ as the quotient of $\mathrm{D}(\mathrm{Bun}_n)$ by a triangulated subcategory $\mathrm{D}_{degen}(\mathrm{Bun}_n)$. (An object $\mathcal{F} \in \mathrm{D}(\mathrm{Bun}_n)$ belongs to $\mathrm{D}_{degen}(\mathrm{Bun}_n)$ essentially when it is degenerate, i.e., when it vanishes in the Whittaker model.)

The quotient $\widetilde{\mathrm{D}}(\mathrm{Bun}_n)$ possesses the following three crucial properties:

1) The perverse t-structure on $\mathrm{D}(\mathrm{Bun}_n)$ induces a well-defined t-structure on $\widetilde{\mathrm{D}}(\mathrm{Bun}_n)$.

2) The Hecke functors $\mathrm{H}_x : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(\mathrm{Bun}_n)$ gives rise to well-defined functors $\widetilde{\mathrm{D}}(\mathrm{Bun}_n) \to \widetilde{\mathrm{D}}(\mathrm{Bun}_n)$, and the latter functors are exact in the sense of the t-structure on $\widetilde{\mathrm{D}}(\mathrm{Bun}_n)$.

3) The subcategory $\mathrm{D}_{degen}(\mathrm{Bun}_n)$ is orthogonal to the subcategory of cuspidal sheaves. I.e., if $\mathcal{F}_1, \mathcal{F}_2 \in \mathrm{D}(\mathrm{Bun}_n)$ are such that $\mathcal{F}_1$ is cuspidal and the image of $\mathcal{F}_2$ in $\widetilde{\mathrm{D}}(\mathrm{Bun}_n)$ vanishes, then $\mathrm{Hom}_{\mathrm{D}(\mathrm{Bun}_n)}(\mathcal{F}_1, \mathcal{F}_2) \simeq \mathrm{Hom}_{\mathrm{D}(\mathrm{Bun}_n)}(\mathcal{F}_2, \mathcal{F}_1) = 0$.

5.4. The main step in the proof that the functor $\mathrm{Av}^d_{n,E}$ is the following:

5.5. Theorem. *The functor* $\mathrm{Av}_{n,E} : \mathrm{D}(\mathrm{Bun}_n) \to \mathrm{D}(\mathrm{Bun}_n)$ *descends to a well-defined functor* $\widetilde{\mathrm{D}}(\mathrm{Bun}_n) \to \widetilde{\mathrm{D}}(\mathrm{Bun}_n)$, *and the latter functor is exact.*

The main idea behind the proof of Theorem 5.5. is the same phenomenon as the one that forbids the existence of Hecke eigen-sheaves on Bun_n with respect to local systems of a wrong rank. Namely, if $\mathcal{F}$ is a perverse sheaf on Bun_n such that the corresponding object of $\mathrm{D}(\mathrm{Bun}_n)$ violates Theorem 5.5., then by looking at the behavior of $\mathrm{H}^{\boxtimes d}(\mathcal{F}) \in \widetilde{\mathrm{D}}(X^d \times \mathrm{Bun}_n)$ around the various diagonals in X^d, we arrive to a contradiction.

Using Theorem 5.5. the proof of the exactness of $\mathrm{Av}^d_{n,E}$ proceeds as follows: From (5.1) (and using the fact that taking invariants is an exact functor), we obtain that the functor $\mathrm{Av}^d_{n,E}$ is well-defined and exact on the quotient category $\widetilde{\mathrm{D}}(\mathrm{Bun}_n)$. Moreover, by induction on n we can assume that for any $\mathcal{F} \in \mathrm{D}(\mathrm{Bun}_n)$, $\mathrm{Av}^d_{n,E}(\mathcal{F})$ is cuspidal.

Hence, if $\mathcal{F}$ is a perverse sheaf on Bun_n, in the cohomological truncations arrows

$$\mathrm{Av}^d_{n,E}(\mathcal{F}) \to \tau^{>0}(\mathrm{Av}^d_{n,E}(\mathcal{F})) \text{ and } \tau^{<0}(\mathrm{Av}^d_{n,E}(\mathcal{F})) \to \mathrm{Av}^d_{n,E}(\mathcal{F}), \qquad (5.2)$$

we have maps between a degenerate and a cuspidal object. However, by property 3 above, both these arrows must be 0. Hence, $\mathrm{Av}^d_{n,E}(\mathcal{F})$ is a perverse sheaf.

References

[BD] A. Beilinson, V. Drinfeld, *Quantization of Hitchin's integrable system and Hecke eigen-sheaves*, Preprint, available at http://www.math.uchicago.edu/~benzvi

[BG] A. Braverman, D. Gaitsgory, *Geometric Eisenstein series*, Preprint math.AG/9912097, to appear in Invent. Math.

[CS] W. Casselman, J. Shalika, *The unramified principal series of p–adic groups II. The Whittaker function*, Comp. Math., **41** (1980), 207–231.

[Dr] V.G. Drinfeld, *Two-dimensional ℓ–adic representations of the fundamental group of a curve over a finite field and automorphic forms on $GL(2)$*, Amer. J. Math., **105** (1983), 85–114.

[FGKV] E. Frenkel, D. Gaitsgory, D. Kazhdan, K. Vilonen, *Geometric realization of Whittaker functions and the Langlands correspondence*, Jour. Amer. Math. Soc., **11** (1998), 451–484.

[FGV] E. Frenkel, D. Gaitsgory, K. Vilonen, *Whittaker patterns in the geometry of moduli spaces of bundles on curves*, Annals of Math., **153** (2001), 699–748.

[FGV1] E. Frenkel, D. Gaitsgory, K. Vilonen, *On the Geometric Langlands conjecture*, Jour. Amer. Math. Soc., **15** (2002), 367–417.

[Ga] D. Gaitsgory, *On a vanishing conjecture appearing in the geometric Langlands correspondence*, Preprint math.AG/0204081 (2002).

[Laf] L. Lafforgue, *Chtoucas de Drinfeld et correspondance de Langlands*, Invent. Math., **147** (2002), 1–241.

[La1] G. Laumon, *Correspondance de Langlands géométrique pour les corps de fonctions*, Duke Math. J., **54** (1987), 309–359.

[La2] G. Laumon, *Faisceaux automorphes pour GL_n: la première construction de Drinfeld*, Preprint alg-geom/9511004 (1995).

[PS] I.I. Piatetski-Shapiro, *Euler subgroups*, in: Lie Groups and Their Representations, ed. I.M. Gelfand, Adam Hilder Publ., (1975), 597–620.

[Sha] J.A. Shalika, *The multiplicity one theorem for GL_n*, Ann. Math., **100** (1974), 171–193.

[Shi] T. Shintani, *On an explicit formula for class 1 Whittaker functions on GL_n over $\mathfrak{P}$–adic fields*, Proc. Japan Acad., **52** (1976), 180–182.

ICM 2002 · Vol. II · 583–597

On the Local Langlands Correspondence

Michael Harris*

Abstract

The local Langlands correspondence for $GL(n)$ of a non-Archimedean local field F parametrizes irreducible admissible representations of $GL(n,F)$ in terms of representations of the Weil-Deligne group WD_F of F. The correspondence, whose existence for p-adic fields was proved in joint work of the author with R. Taylor, and then more simply by G. Henniart, is characterized by its preservation of salient properties of the two classes of representations.

The article reviews the strategies of the two proofs. Both the author's proof with Taylor and Henniart's proof are global and rely ultimately on an understanding of the ℓ-adic cohomology of a family of Shimura varieties closely related to $GL(n)$. The author's proof with Taylor provides models of the correspondence in the cohomology of deformation spaces, introduced by Drinfeld, of certain p-divisible groups with level structure.

The general local Langlands correspondence replaces $GL(n,F)$ by an arbitrary reductive group G over F, whose representations are conjecturally grouped in packets parametrized by homomorphisms from WD_F to the Langlands dual group ${}^LG$. The article describes partial results in this direction for certain classical groups G, due to Jiang-Soudry and Fargues.

The bulk of the article is devoted to motivating problems that remain open even for $GL(n)$. Foremost among them is the search for a purely local proof of the correspondence, especially the relation between the Galois-theoretic parametrization of representations of $GL(n,F)$ and the group-theoretic parametrization in terms of Bushnell-Kutzko types. Other open questions include the fine structure of the cohomological realization of the local Langlands correspondence: does the modular local Langlands correspondence of Vigneras admit a cohomological realization?

2000 Mathematics Subject Classification: 11, 14, 22.

Introduction

Compared to the absolute Galois group of a number field, e.g. $Gal(\overline{\mathbb{Q}}/\mathbb{Q})$, the Galois group Γ_F of a non-archimedean local field F has a ridiculously simple structure. Modulo the inertia group I_F, there is a natural isomorphism

$$\Gamma_F/I_F \xrightarrow{\sim} Gal(\bar{k}_F/k_F),$$

*Institut de Mathématiques de Jussieu-UMR CNRS 7586, Université Paris 7. Membre, Institut Universitaire de France, France. E-mail: harris@math.jussieu.fr

where k_F is the residue field of F. Then $Gal(\bar{k}_F/k_F)$ is topologically generated by the geometric Frobenius $Frob(x) = x^{\frac{1}{q}}$, where $q = |k_F| = p^f$ for p prime. The inertia group has a two step filtration,

$$1 \to P_F \to I_F \to \prod_{\ell \neq p} \mathbb{Z}_\ell \to 1,$$

where the wild ramification group P_F is a pro-p group.

Thus if $\sigma : \Gamma_F \to GL(n, \mathbb{C})$ is a continuous homomorphism, $n \geq 1$, then the image of σ is solvable, and $\sigma(P_F)$ is nilpotent. This is still true when σ is a finite-dimensional complex representation of the *Weil group*, the subgroup $W_F \subset \Gamma_F$ of elements whose image in $Gal(\bar{k}_F/k_F)$ is an integral power of $Frob$. Despite this simplicity, our understanding of the set of equivalence classes of n-dimensional representations of W_F is far from complete, at least when p divides n.

The reciprocity map of local class field theory:

$$F^\times \xrightarrow{\sim} W_F^{ab},$$

identifies the set $\mathcal{G}(1, F)$ of one-dimensional representations of W_F with the set $\mathcal{A}(1, F)$ of irreducible representations of $F^\times = GL(1, F)$. More than a simple bijection, this identification respects a number of salient structures, and its behavior with respect to field extensions F'/F is well understood. Moreover, it is compatible, in a straightforward way, with global class field theory, and was historically first derived as a consequence of the latter.

A simple special case of Langlands' functoriality principle is the so-called strong Artin conjecture, which identifies the Artin L-function attached to an irreducible n-dimensional representation of $Gal(\overline{\mathbb{Q}}/K)$, for a number field K, as the L-function of a cuspidal automorphic representation of $GL(n)_K$. As a local counterpart, Langlands proposed a parametrization of irreducible admissible representations of reductive groups over the local field F in terms of representations of W_F. The prototypical example is the local Langlands conjecture for $GL(n)$. By analogy with the case $n = 1$, the set of equivalence classes of irreducible admissible representations of $G_n = GL(n, F)$ is denoted $\mathcal{A}(n, F)$. By $\mathcal{G}(n, F)$ we denote the set of equivalence classes of n-dimensional representations, not of W_F but rather of the Weil-Deligne group WD_F, and only consider representations for which any lifting of $Frob$ acts semisimply. Then the general local Langlands conjecture for $GL(n)$, in its crudest form, asserts the existence of a family of bijections, as F and n vary:

$$\sigma = \sigma_{n,F} : \mathcal{A}(n, F) \xrightarrow{\sim} \mathcal{G}(n, F). \tag{0.1}$$

A normalization condition is that the central character ξ_π of $\pi \in \mathcal{A}(n, F)$ correspond to $\det \sigma(\pi)$ via local class field theory.

The first general result of this type was proved by Henniart [He1]. Early work of Bernstein and Zelevinsky reduced (0.1) to the existence of bijections

$$\sigma = \sigma_{n,F} : \mathcal{A}_0(n, F) \xrightarrow{\sim} \mathcal{G}_0(n, F), \tag{0.2}$$

where $\mathcal{G}_0(n,F)$ are the irreducible representations of W_F and $\mathcal{A}_0(n,F)$ is the supercuspidal subset of $\mathcal{A}(n,F)$. Both sides of (0.2) are homogeneous spaces under $\mathcal{A}(1,F)$, and thus under its subset $\mathcal{A}^{unr}(1,F)$ of unramified characters χ of $F^\times$: if $\pi \in \mathcal{A}_0(n,F)$ (resp. $\sigma \in \mathcal{G}_0(n,F)$), we denote by $\pi \otimes \chi$ (resp. $\sigma \otimes \chi$) the tensor product of π (resp. σ) with the one-dimensional representation $\chi \circ \det$ of G_n (resp. with the character $\sigma_{1,F}(\chi)$ of W_F). Each $\mathcal{A}^{unr}(1,F)$-orbit on either side of (0.2) has a discrete invariant, the Artin conductor $a(\pi)$, resp. $a(\sigma)$, and the sets $\mathcal{A}_0(n,F)[a]$, resp. $\mathcal{G}_0(n,F)[a]$ of orbits with given Artin conductor a are known to be finite. The main theorem of [He1] is the *numerical local Langlands correspondence*

$$|\mathcal{A}_0(n,F)[a]| = |\mathcal{G}_0(n,F)[a]|, \tag{0.3}$$

established by painstakingly counting both sides.

It has been known for some time that a family of bijections (0.2), compatible with Artin conductors and twists by $\mathcal{A}(1,F)$, is not unique. Henniart showed (the Uniqueness Theorem, [He2]) that at most one normalized bijection is compatible with contragredients and twists and satisfies the condition:

$$L(s,\pi\otimes\pi') = L(s,\sigma(\pi)\otimes\sigma(\pi')); \qquad \varepsilon(s,\pi\otimes\pi',\psi) = \varepsilon(s,\sigma(\pi)\otimes\sigma(\pi'),\psi) \tag{0.4}$$

for $\pi \in \mathcal{A}_0(n,F)$, $\pi' \in \mathcal{A}_0(n',F)$, $n' < n$. Here $\psi : F \to \mathbb{C}^\times$ is a non-trivial character. The L- and ε-factors are defined on the automorphic side in [JPS, Sh]; on the Galois side by Langlands and Deligne (cf. [D]). It is in this version that the local Langlands conjecture for $GL(n)$ has finally been established: for fields of positive characteristic in [LRS], and for p-adic fields in [HT], followed shortly thereafter by [He3] (see also [He5]).

1. Compatibility with global correspondences

As in the first proofs of local class field theory, the bijections (0.2) are constructed in [LRS,HT,He3] by local specialization of maps for certain global fields E:

$$\sigma = \sigma_{n,E} : \mathcal{A}^{good}(n,E) \hookrightarrow \mathcal{G}(n,E). \tag{1.1}$$

Here E is supposed to have a place w such that $E_w \xrightarrow{\sim} F$, $\mathcal{A}^{good}(n,E)$ is a class of cuspidal automorphic representations of $GL(n)_E$ chosen to fit the circumstances, and $\mathcal{G}(n,E)$ can be taken to be the set of equivalence classes of compatible families of n-dimensional semi-simple λ-adic representations of $Gal(\bar{E}/E)$. In particular, both sides of (1.1) as well as (0.2) are taken with ℓ-adic, rather than complex, coefficients; this does not change the problem in an essential way.

The map σ of (0.1) is particularly simple for unramified representations. An unramified $\tau \in \mathcal{G}(n,F)$ is given by an unordered n-tuple $(\chi_1,\ldots,\chi_n)$ of unramified characters of $W_F^{ab} \xrightarrow{\sim} F^\times$. Ordering the χ_i arbitrarily, we obtain a character χ of the Levi subgroup G_1^n of a Borel subgroup $B \subset G_n$. The element of $\mathcal{A}(n,F)$

corresponding to τ is then the unique subquotient $\pi(\tau) = \sigma^{-1}(\tau)$ of the normalized induced representation $Ind_B^{GL(n,F)}\chi$ containing a vector fixed by $GL(n, \mathcal{O}_F)$, $\mathcal{O}_F$ the integer ring of F. This defines a bijection, a special case of the *Satake parametrization*, between the unramified subset $\mathcal{G}^{unr}(n, F)$ and the unramified (spherical) representations $\mathcal{A}^{unr}(n, F)$ of G_n.

Fix an automorphic representation $\Pi = \otimes_v \Pi_v$ of $GL(n)_E$. The representation $\sigma_{n,E}(\Pi)$, when it exists, should have the property that

$$\sigma_{n,E}(\Pi) \mid_{W_{E_v}} = \sigma_{n,E_v}(\Pi_v) \tag{1.2}$$

for almost all v such that $\Pi_v \in \mathcal{A}^{unr}(n, E_v)$; i.e., all but finitely many v. By Chebotarev density, this determines $\sigma_{n,E}(\Pi)$ uniquely. One can then hope that

Hope 1.3. *$\sigma_{n,E}(\Pi)_{W_{E_v}}$ depends only on F and Π_v for all v,*

including $v = w$, the place of interest. Setting $\sigma_{n,F}(\Pi_v) = \sigma_{n,E}(\Pi)_{W_{E_v}}$, one then needs to show that

1.4. *For any $\pi \in \mathcal{A}_0(n, F)$ there exists $\Pi \in \mathcal{A}^{good}(n, E)$, for some E, with $\Pi_w \simeq \pi$;*

1.5. *For $\Pi \in \mathcal{A}^{good}(n, E)$, $\Pi' \in \mathcal{A}^{good}(n', E)$, the completed L-function $\Lambda(s, \sigma_{n,E}(\Pi) \otimes \sigma_{n',E}(\Pi'))$ satisfies the functional equation*

$$\Lambda(s, \sigma_{n,E}(\Pi) \otimes \sigma_{n',E}(\Pi')) = \varepsilon(s, \sigma_{n,E}(\Pi) \otimes \sigma_{n',E}(\Pi'))\Lambda(1 - s, \check{\sigma}_{n,E}(\Pi) \otimes \check{\sigma}_{n',E}(\Pi'));$$

$$\varepsilon(s, \sigma_{n,E}(\Pi) \otimes \sigma_{n',E}(\Pi')) = \prod_v \varepsilon_v(s, \sigma_{n,E}(\Pi) \otimes \sigma_{n',E}(\Pi'), \psi_v)$$

is the product of local Deligne-Langlands ε factors.

Here $\check{}$ denotes contragredient. The local additive characters ψ_v are assumed to be the local components of a continuous character of $\mathbf{A}_E/E$.

1.6. *The map $\sigma = \sigma_{n,F} : \mathcal{A}_0(n, F) \to \mathcal{G}(n, F)$*

(i) *takes values in $\mathcal{G}_0(n, F)$;*
(ii) *defines a bijection $\mathcal{A}_0(n, F) \leftrightarrow \mathcal{G}_0(n, F)$;*
(iii) *satisfies the remaining requirements of a local Langlands correspondence, especially (0.4).*

The main burden of [LRS] is the construction of a class $\mathcal{A}^{good}(n, E)$ large enough to satisfy (1.4): now a moot point, since Lafforgue has proved that all cuspidal automorphic representations of $GL(n)$ of a function field are "good" in this sense. The $\mathcal{A}^{good}(n, E)$ in [LRS] are the automorphic representations that contribute to the cohomology of an appropriate Drinfeld modular variety, constructed from scratch for the occasion, attached to the multiplicative group of a division algebra of dimension n^2 over E, unramified at the chosen w. Property (1.5) in this

case follows from general results of Deligne in [D], valid only in equal characteristic. Now by (1.2), for a sufficiently large set S of places of E we have

$$\prod_{v \notin S} L(s, \Pi_v \times \Pi'_v) = \prod_{v \notin S} L(s, \sigma_{n,E_v}(\Pi_v) \otimes \sigma_{n',E_v}(\Pi'_v)), \tag{1.7}$$

where the left-hand side is the Rankin-Selberg L-function. Completing the latter to $\Lambda(s, \Pi \otimes \Pi')$ and applying [JPS] or [Sh], we find the functional equation

$$\Lambda(s, \Pi \otimes \Pi') = \prod_v \varepsilon_v(s, \Pi \otimes \Pi', \psi_v)\Lambda(1-s, \check{\Pi} \otimes \check{\Pi}'). \tag{1.8}$$

In other words, the partial L-functions, identified via (1.8), satisfy *two* functional equations (1.5) and (1.8). An argument first used by Henniart then yields (0.4), and then (1.3) and the full local Langlands conjecture follow from the Uniqueness Theorem of [He2].

When F is p-adic a class $\mathcal{A}^{CK}(n, E)$ satisfying (1.2) is implicitly defined by work of Clozel and Kottwitz [K,Cl1], provided E is a CM field. For $\mathcal{A}^{CK}(n, E)$ one can take cuspidal automorphic representations Π, cohomological at all archimedean primes, square integrable at several finite primes other than w, and such that $\check{\Pi} \simeq \Pi^c$, where c denotes conjugation of E over its maximal totally real subfield. However, the Galois-theoretic functional equation (1.5) is only available a priori when $\sigma_{n,E}(\Pi)$ is associated to a global complex representation of the Weil group of E; i.e. when $\sigma_{n,E}(\Pi)$ becomes abelian over a finite extension of E. The article [H2] showed that there were enough Π of this type in $\mathcal{A}^{CK}(n, E)$. Denoting by $\mathcal{A}^{good}(n, E)$ the set of such Π, we find that (1.4) is impossible as soon as p divides n; however, an argument in [H2], based on Brauer's theorem on induced characters and (0.3), shows that (1.4) is true "virtually," in the set of formal sums with integral coefficients of elements of $\mathcal{A}^{good}(n, E)$ for varying n. It then suffices to prove the following weak form of (1.3), which occupies the bulk of [HT]:

Theorem 1.9 [HT]. *For all $\Pi \in \mathcal{A}^{CK}(n, E)$, the semisimplification $\sigma_{n,E}(\Pi)_{W_{E_v},ss}$ of $\sigma_{n,E}(\Pi)_{W_{E_v}}$ depends only on F and Π_v for all v.*

More precisely, [HT] proves that $\sigma_{n,E}(\Pi)_{W_F,ss}$ can be calculated explicitly in the vanishing cycles of certain formal deformation spaces $M^h_{LT,F}$ defined by Drinfeld (see §2). Following [K,Cl1], the representations $\sigma_{n,E}(\Pi)$ are initially realized in the cohomology of certain Shimura varieties with canonical models over E, and (1.9) is proved by a study of their bad reduction at w. Henniart soon realized that, for $\Pi \in \mathcal{A}^{good}(n, E)$, the purely local nature of $\sigma_{n,E}(\Pi)_{W_{E_v},ss}$, and hence the definition of a map $\sigma_{n,F}$, could be derived directly from (1.5) and from the results of [He1,He2]. Though [He3] dispenses with the geometry, it is still a global argument inasmuch as it relies on [H2], which in turn depends on [K,Cl1] and [CL].the conditional base change results of [CL].

A global consequence of Theorem 1.9 is the *Generalized Ramanujan Conjecture* for the automorphic representations in $\mathcal{A}^{CK}(n, E)$: if $\Pi \in \mathcal{A}^{CK}(n, E)$ and is unitary

then its local component Π_v is tempered at every finite prime v. Clozel in [Cl1] already showed this to be true for almost all unramified v. Generalizing a method developed by Lubotzky, Phillips, and Sarnak for the 2-sphere, Clozel [Cl2] uses the version of the Generalized Ramanujan Conjecture proved in [HT] to obtain effective constructions of families of equidistributed points on odd-dimensional spheres.

With (0.1) out of the way, we can propose the following improvement of (1.3):

Problem 1. Show that

$$\sigma_{n,E}(\Pi)_{W_{E_v}} \xrightarrow{\sim} \sigma_{n,E_v}(\Pi_v). \tag{1.10}$$

For $n = 2$ this was established by Carayol assuming standard conjectures on the semisimplicity of Frobenius. Theorem 1.9 shows that it holds up to semisimplification. The techniques of [HT], like the earlier work of Kottwitz treating unramified places, is based on a comparison of trace formulas, and cannot detect the difference between two representations with the same semisimplification. Assuming semisimplicity of Frobenius, the equality (1.10) follows easily from Theorem 1.9 and Deligne's conjecture, apparently inaccessible, on the purity of the monodromy weight filtration.

Compatibility with functoriality. Given cuspidal automorphic representations Π_i of $GL(n_i)_E$, for $i = 1, 2, \ldots, r$, and a homomorphism $\rho : GL(n_1) \times \cdots \times GL(n_r) \to GL(N)$ of algebraic groups, Langlands functoriality predicts the existence of an automorphic representation $\rho_*(\Pi_1 \otimes \ldots \Pi_r)$, not necessarily cuspidal, of $GL(N)_E$, such that, for almost all unramified places v of E,

$$\sigma_{N,E_v}(\rho_*(\Pi_1 \otimes \ldots \Pi_r)_v) = \rho \circ (\otimes_{i=1}^r \sigma_{n_i,E_v}(\Pi_{i,v})). \tag{1.11}$$

In recent years this has been proved for general number fields E in several important special cases: the tensor products $GL(2) \times GL(2) \to GL(4)$ (Ramakrishnan) and $GL(2) \times GL(3) \to GL(6)$ (Kim-Shahidi), and the symmetric powers $Sym^3 : GL(2) \to GL(4)$ (Kim-Shahidi) and $Sym^4 : GL(2) \to GL(5)$ (Kim). It has been verified in all four cases that (1.11) holds for all v.

Construction of supercuspidal representations by "backwards lifting".

The unitary representation $\pi \in \mathcal{A}_0(n, F)$ is isomorphic to its contragredient if and only if the local factor $L(s, \pi \times \pi)$ has a pole at $s = 0$, which is necessarily simple. The local factor can be decomposed as a product:

$$L(s, \pi \times \pi) = L(s, \pi, Sym^2) L(s, \pi, \wedge^2), \tag{1.12}$$

where the two terms on the right are defined for unramified π by Langlands and in general by Shahidi. Only one of the factors on the right has a pole. Using the class $\mathcal{A}^{good}(n, E)$ of automorphic representations, Henniart has shown that it is the first factor (resp. the second factor) if and only if $\sigma(\pi)$ is orthogonal (resp. symplectic);

the symplectic case only arises for n even. One thus expects that π is obtained by functorial transfer from an L-packet of a classical group G over F, via the map of L-groups ${}^LG \to GL(n,\mathbb{C})$, where ${}^LG = SO(n,\mathbb{C})$, resp. $Sp(n,\mathbb{C})$, if the first, resp. the second factor in (1.12) has a pole at $s = 0$.

In particular, when $n = 2m$ and $L(s, \pi, \wedge^2)$ has a pole at $s = 0$, π should come from an L-packet on the split group $SO(2m+1, F)$. Using a local analogue of the method of "backwards lifting," or automorphic descent, due to Ginzburg, Rallis, and Soudry, Jiang and Soudry have constructed a generic supercuspidal representation π' of $SO(2m+1, F)$ for every $\pi \in \mathcal{A}_0(n, F)$ with the indicated pole. More generally, they have obtained a complete parametrization of generic representations of split $G = SO(2m+1, F)$ in terms of Langlands parameters $WD_F \to {}^LG$ [JS]. These results should certainly extend to other classical groups.

2. Cohomological realizations of the local correspondence

The theory of the new vector implies easily that any irreducible admissible representation $\pi \in \mathcal{A}(n, F)$ has a rational model over the field of definition of its isomorphism class: the Brauer obstruction is trivial for G_n. The analogous assertion fails for representations in $\mathcal{G}(n, F)$. Thus one cannot expect the existence of a space $\mathcal{M}$, with a natural action of $G_n \times W_F$, whose cohomology of whatever sort realizes the local Langlands correspondence, as an identity of virtual representations

$$\sigma_{n,F}(\pi) = \pm[Hom_{G_n}(H_c(\mathcal{M}), \pi)] := \pm \sum_i (-1)^i Hom_{G_n}(H_c^i(\mathcal{M}), \pi). \tag{2.1}$$

We add a third group to the picture by taking J to be an inner form of G_n, the multiplicative group of a central simple algebra D over F of dimension n^2, with Hasse invariant $\frac{r_D}{n}$. The set $\mathcal{A}(n, F)$ contains a subset $\mathcal{A}_{(2)}(n, F)$ of discrete series representations, character twists of those realized in the regular representation on $L_2(G_n)$ (modulo center). The set $\mathcal{A}(J)$ of equivalence classes of irreducible admissible representations contains an analogous subset $\mathcal{A}_{(2)}(J)$, equal to $\mathcal{A}(J)$ if D is a division algebra. The *Jacquet-Langlands correspondence* [R,DKV] is a bijection $JL : \mathcal{A}_{(2)}(G_n) \xrightarrow{\sim} \mathcal{A}_{(2)}(J)$ determined by the identity of distribution characters

$$\chi_\pi(g) = \varepsilon(J)\chi_{JL(\pi)}(j), \ \pi \in \mathcal{A}_{(2)}(G) \tag{2.2}$$

if $\varepsilon(J) = \pm 1$ is the Kottwitz sign and g and j are elliptic regular elements with the same eigenvalues. When $r_D = 1$ there are two spaces $\tilde{\Omega}_F^n$ and $\mathcal{M}_{LT,F}^n$ with natural $G_n \times J$-actions. The former is a countable union, indexed by $\mathbb{Z}$, of copies of the profinite étale cover $\tilde{\Omega}_F^{n,0}$ of the rigid-analytic upper half space $\Omega_F^n = \mathbb{P}^{n-1}(\mathbb{C}_p) - \mathbb{P}^{n-1}(F)$, defined by Drinfeld in [D2]. The latter is the rigid generic fiber of the formal deformation space $M_{LT,F}^n$ of the one-dimensional height n formal $\mathcal{O}_F$-module with Drinfeld level structures of all degrees [D1]. A relation analogous to (2.1) was conjectured by Carayol in [C1], with $\pm = (-1)^{n-1}$:

Theorem 2.3. *For π supercuspidal*

$$\sigma^{\#}(\pi) \otimes JL(\pi) = \pm[Hom_{G_n}(H_c(\tilde{\Omega}^n_F), \pi)]$$
$$\sigma^{\#}(\pi) \otimes \pi = \pm[Hom_J(H_c(\mathcal{M}^n_{LT,F}), JL(\pi))].$$

The notation $\sigma^{\#}(\pi)$ indicates that $\sigma(\pi)$ has been twisted by an elementary factor. We use the rigid-analytic étale cohomology introduced by Berkovich [B] with coefficients in $\overline{\mathbb{Q}}_\ell$, $\ell \neq p$. For $\mathcal{M}^n_{LT,F}$ this can be interpreted as a space of vanishing cycles for the formal deformation space, viewed as a formal scheme over $Spf(\mathcal{O}_F)$. The case of $\tilde{\Omega}^n_F$ was proved in [H1], using the existence of Shimura varieties admitting rigid-analytic uniformizations by $\tilde{\Omega}^n_F$. This has recently been extended to F of equal characteristic by Hausberger [Hau]. The case of $\mathcal{M}^n_{LT,F}$, again for π supercuspidal, was initially treated by Boyer [Bo] in the equal-characteristic case. The analogous statement for F p-adic, and for any π, is the logical starting point of the proof of Theorem 1.9 in [HT].

Theorem 2.3 is extended in [HT] to general $\pi \in \mathcal{A}_{(2)}(G)$. The explicit formula for the alternating sum of the $Hom_J(H^i_c(\mathcal{M}^n_{LT,F}), JL(\pi))$ is awkward but yields a simple expression for

$$\sum_{i,j}(-1)^{i+j} Ext^j_G(Hom_J(H^i_c(\mathcal{M}^n_{LT,F}), JL(\pi)), \pi)$$
$$= \sum_{i,j,k}(-1)^{i+j+k} Ext^j_G(Ext^k_J(H^i_c(\mathcal{M}^n_{LT,F}), JL(\pi)), \pi) \quad (2.4)$$

in terms of the semisimplification of $\sigma(\pi)$. An analogous *conjectural* expression for *individual* $H^i_c(\tilde{\Omega}_F)$ has been circulating for several years and should appear in a forthcoming joint paper with Labesse. Faltings has proved [F2] that the spaces $\tilde{\Omega}_F$ and $\mathcal{M}^n_{LT,F}$ become isomorphic after p-adic completion of the latter. Thus the two questions in the following problem reduce to a single question:

Problem 2. Determine the individual representations $H^i_c(\mathcal{M})$, and the spaces $Ext^j_{G_n}(H^i_c(\tilde{\Omega}_F), \pi)$ and $Hom_J(H^i_c(\mathcal{M}^n_{LT,F}), JL(\pi))$ for all i, j, all $\pi \in \mathcal{A}(n, F)$. In particular, show that $Ext^j_{G_n}(H^i_c(\tilde{\Omega}_F), \pi)$ vanishes unless there exists $\pi' \in \mathcal{A}_{(2)}(n, F)$ such that π and π' induce the same character of the Bernstein center.

The results of [HT] imply that, for any $\pi \in \mathcal{A}_{(2)}(n, F)$, with Bernstein character β_π, the Bernstein center acts on $\sum_i (-1)^i Hom_J(H^i_c(\mathcal{M}^n_{LT,F}), JL(\pi))$ via β_π.

For π supercuspidal it is known in all cases that the spaces in Problem 2 vanish for $i \neq n-1$ (and for $j \neq 0$). This vanishing property should characterize supercuspidal π among representations in $\mathcal{A}_{(2)}(n, F)$. When π is the Steinberg representation, the $H^i_c(\Omega_F)$, as well as the corresponding Ext groups, are calculated explicitly in [SS]. The calculation in [SS] is purely local, whereas the vanishing outside the middle degree for $\pi \in \mathcal{A}_0(n, F)$ is based on properties of automorphic forms.

Problem 3. Find a purely local proof of the vanishing property for $\pi \in \mathcal{A}_0(n, F)$.

The covering group of $\tilde{\Omega}_F^{n,0}$ over Ω_F^n can be identified with the maximal compact subgroup $J^0 \subset J$. Thus $H_c^i(\tilde{\Omega}_F)$ can be written as a sum $\oplus_\tau H_c^i(\tilde{\Omega}_F)[\tau]$ of its τ-isotypic components, where τ runs over irreducible representations of J^0 or, equivalently, over inertial equivalence classes of representations of J. Closely related to Problem 3 is the following

Problem 4. Characterize $\tau \in \mathcal{A}(J)$ such that $JL^{-1}(\tau) \in \mathcal{A}_0(n, F)$. Equivalently, calculate the Jacquet functors of the G_n-spaces $H_c^i(\tilde{\Omega}_F)[\tau]$ geometrically, in terms of τ.

When n is prime $JL^{-1}(\tau) \in \mathcal{A}_0(n, F)$ if and only if $\dim \tau > 1$; when $\dim \tau = 1$ $JL^{-1}(\tau)$ is a twist of the Steinberg representation. For general n practically nothing is known.

Results of L. Fargues [Fa]. For certain classical F-groups G, Rapoport and Zink, using the deformation theory of p-divisible (Barsotti-Tate) groups, have defined pro-rigid analytic spaces $\mathcal{M}$ admitting continuous $G \times J \times W_E$ actions on their ℓ-adic cohomology, where J is an inner form of G and E, the *reflex field* of $\mathcal{M}$, is a finite extension of F [RZ]. In [R] Rapoport proposes a conjectural formula, which he attributes to Kottwitz, for the discrete series contribution to the virtual $G \times J \times W_E$-module $[H(\mathcal{M})] = \sum_i (-1)^i H_c^i(\mathcal{M}, \overline{\mathbb{Q}}_\ell)$. The pairs (G, J) considered in [RZ] include $(G_n, D^\times)$ with general Hasse invariant $\frac{r_D}{n}$, $G = J = GU(n)$, the quasi-split unitary similitude group attached to the unramified quadratic extension of F, and the symplectic similitude group $G = GSp(2n, F)$.

Theorem 2.5 (Fargues). *Suppose $F/\mathbb{Q}_p$ unramified, $(G, J) = (G_n, D^\times)$, with $(r_D, n) = 1$. For any $\pi \in \mathcal{A}_0(n, F)$ we have*

$$\sum_i (-1)^i Hom_J(H_c^i(\mathcal{M}, \overline{\mathbb{Q}}_\ell), JL(\pi))_0 = \pm \pi \otimes \wedge^? \sigma(\pi)$$

up to a simple twist. Here the subscript $_0$ denotes the G-supercuspidal part and $\wedge^?$ is a certain tensor product of exterior powers of $\sigma(\pi)$ with total weight r_D, depending on auxiliary data defining $\mathcal{M}$.

This confirms the Kottwitz-Rapoport conjectures in the case in question. For $G = J = GU(3)$ Rogawski has defined a local Langlands correspondence via base change to $GL(3)$. In that case the supercuspidal representations of G are grouped into L-packets. Fargues's techniques apply to this case as well, and he obtains a version of the Kottwitz-Rapoport conjectures, more difficult to state than Theorem 2.5 (higher *Ext*'s are involved, and the formula is averaged over L-packets).[1] More

[1]The statement of the general Kottwitz-Rapoport conjectures in [H3] for general discrete series representations is based on a misreading of Rapoport's use of the term "discrete L parameter". The correct conjecture should involve the analogue of the alternating sum on the right-hand side of (2.4), with $JL(\pi)$ replaced by π' in the L-packet associated to π.

generally, Fargues' methods apply to classical groups attached to Shimura varieties, whenever the trace formula is known to be stable and functorial transfer from G to $GL(n)$ has been established.

In contrast to [HT], Fargues' methods are essentially rigid-analytic, and make no use of equivariant regular integral models of Shimura varieties in wildly ramified level – fortunately so, since such models are not known to exist. Heuristically, the characters of the representations of G and J on $[H(\mathcal{M})]$ can be related by applying a Lefschetz trace formula to ℓ-adic cohomology of the rigid space $\mathcal{M}$. This approach, which in principle provides no information about the W_F action, has been successfully applied to $\tilde{\Omega}^n_F$ by Faltings in [F1], and to $M^n_{LT,F}$ by Strauch [S] when $n = 2$. For higher $M^n_{LT,F}$, and for the Rapoport-Zink spaces studied by Fargues, one does not yet know how to deal with wild boundary terms in Huber's Lefschetz formula [Hu] and its higher-dimensional generalizations.

Using work of Oort and Zink on stratification of families of abelian varieties and the slope filtration for p-divisible groups, Mantovan [M] has developed another approach to the cohomology of Shimura varieties of PEL type. Closer in spirit to [HT] than to [F], [M] obtains finer results on the geometry of the special fiber and a description of the cohomology in ramified level similar to that of [F].

Cohomological realizations with torsion coefficients.

It would be convenient if the following question had an affirmative answer:

Question 5. Is $H^i_c(\tilde{\Omega}^n_F, \mathbb{Z}_\ell)$ a torsion-free $\mathbb{Z}_\ell$-module?

The global trace formula methods used in [H1] and [HT] to derive Theorem 2.3 from an analysis of the cohomology of the "simple" Shimura varieties of the title of [K] are insensitive to torsion in cohomology. When $\ell > n$ it may be possible, as in recent work of Mokrane and Tilouine, to combine ℓ-adic Hodge theory with the generalized Eichler-Shimura congruence formula, for the same "simple" Shimura varieties, to answer Question 5. For $\ell \leq n$ completely new ideas are needed.

When k is an algebraically closed field of characteristic $\ell \neq p$, Vignéras has defined a class of smooth supercuspidal representations $\mathcal{A}_{0,k}(n, F)$ of G_n with coefficients in k, and has proved that they are in bijection with the set $\mathcal{G}_{0,k}(n, F)$ of irreducible n-dimensional representations of W_F over k (see article in these Proceedings). It is natural to expect that this modular local Langlands correspondence is realized on the spaces $H^\bullet_c(\mathcal{M}, k)$, with $\mathcal{M} = \tilde{\Omega}^n_F$ or $\mathcal{M}^n_{LT,F}$.

Problem 6. Define a modular Jacquet-Langlands map $\pi \mapsto JL(\pi)$ from $\mathcal{A}_{0,k}(n, F)$ to k-representations of J, and formulate the last sentence precisely. Does the virtual W_F-module

$$(-1)^{n-1} \sum_{i,j,k} (-1)^{i+j+k} Ext^j_G(Ext^k_J(H^i_c(\mathcal{M}, k), JL(\pi)), \pi)$$

realize the modular local Langlands correspondence?

Implicit in the second question is the assumption that the modular Jacquet-Langlands map can be extended to a wider class of k-representations of G_n, perhaps including reduction (mod ℓ) of supercuspidal representations in characteristic zero. One can of course ask the same questions when $\ell = p$. In this case we can consider rigid (de Rham) cohomology, in the sense of Berthelot, as well as p-adic étale cohomology. All three groups G_n, J, and W_F have large analytic families of p-adic representations. It is not at all clear whether the p-adic cohomology of $\tilde{\Omega}^n_F$ is sufficiently rich to account for all p-adic deformations – in categories yet to be defined – of a given representation occurring in cohomology with coefficients in $\overline{\mathbb{F}}_p$.

3. Explicit parametrization of supercuspidal representations

Distribution characters.

The distribution character χ_π, a locally integrable function on the set of regular semisimple elements of $G_n = GL(n, F)$, is the fundamental analytic invariant of $\pi \in \mathcal{A}(n, F)$. For $\pi \in \mathcal{A}_{(2)}(n, F)$, $\chi_{JL(\pi)}$, related to χ_π by (2.2), extends continuously to an invariant function on $J = D^\times$ provided $(r_D, n) = 1$, which we assume. Under this hypothesis every element of J is elliptic and every elliptic regular element j is contained in a unique maximal torus $T(j)$, isomorphic to the multiplicative group of an extension K of F of degree n. Since $JL(\pi)$ is finite-dimensional, its restriction to $T(j)$ equals $\sum_\xi a_\pi(\xi)\xi$ where ξ runs over characters of $K_j^\times$ and the coefficients $a_\pi(\xi) = a_\pi(K, \xi)$ are non-negative integers, almost all zero. In this way $\pi \in \mathcal{A}_{(2)}(n, F)$ is determined by the integer-valued function $a_\pi(K, \xi)$ where K runs over degree n extensions of F and ξ over characters of $K^\times$. Invariance entails the symmetry condition $a_\pi(K', {}^\sigma\xi) = a_\pi(K, \xi)$ where $\sigma : K \xrightarrow{\sim} K'$ is an isomorphism over F; in particular, if $\sigma \in Aut_F(K)$.

Problem 7. Express $a_\pi(K, \xi)$ in terms of numerical invariants of $\sigma(\pi)$.

Of course $a_\pi(K, \xi) = 0$ unless $\xi|_{F^\times}$ coincides with the central character ξ_π of π. When $n = 2$ $a_\pi(K, \xi) \in \{0, 1\}$, and a theorem of Tunnell, completed by H. Saito, relates the nonvanishing of $a_\pi(K, \xi)$ to the local constant $\varepsilon(\frac{1}{2}, \sigma(\pi) \otimes \xi^{-1}, \psi)$. For n prime to p a conjecture of Reimann, following an earlier conjecture of Moy, expresses χ_π in terms of $\sigma(\pi)$; work in progress of Bushnell and Henniart shows that this conjecture is almost right (probably up to an unramified character of degree at two).

Parametrization via types.

A fundamental theorem of Bushnell and Kutzko asserts that every supercuspidal π can be obtained by compactly supported induction from a finite-dimensional representation τ of a subgroup $H \subset G_n$ which is compact modulo the center Z_n of G_n. The pair (H, ρ), called an *extended type*, is unique up to conjugation by G_n. The character χ_π can be obtained from (H, ρ) by a simple integral formula [BH, (A.14)].

The outstanding open problem concerning the local Langlands correspondence is undoubtedly

Problem 8. (a) Define $\sigma(\pi)$ directly in terms of (H, ρ) (and vice versa).

(b) Show directly that the definition of σ in (a) has the properties of a local Langlands correspondence.

Note that (b) presupposes a direct construction of the local Galois constants.

Problem 8 formulates the hope, often expressed, for a purely local construction of the local Langlands correspondence. Bushnell, Henniart, and Kutzko have made considerable progress toward this goal. Among other results, they have obtained:

- A formula for the conductor $a(\pi \times \pi')$, $\pi \in \mathcal{A}_0(n, F)$, $\pi' \in \mathcal{A}_0(n', F)$ [BHK];
- A purely local candidate for the base change map $\mathcal{A}(n, F) \to \mathcal{A}(n, K)$ when K/F is a tame, not necessarily Galois extension [BH, I], agreeing with Arthur-Clozel base change for K/F cyclic;
- A bijection between wildly ramified supercuspidal representations of G_{p^m} and wildly ramified[2] representations in $\mathcal{G}_0(p^m, F)$, preserving local constants [BH, II].

In each instance, the constructions and proofs are based primarily on the theory of types. A complete solution of Problem 8 remains elusive, however, absent a better understanding of the local Galois constants.

Question 9. Can the types (H, ρ) be realized in the cohomology (ℓ-adic or p-adic) of appropriate analytic subspaces of $\tilde{\Omega}^n_F$ or $\mathcal{M}^n_{LT,F}$?

Positive results for certain (H, ρ) have been announced by Genestier and Strauch, at least when $n = 2$.

Acknowledgments. I thank R. Taylor, G. Henniart, and L. Fargues for their comments on earlier versions of this report.

References

More or less detailed accounts of the history of the local Langlands conjecture, and of its proofs, can already be found in the literature: [Rd] and [Ku] describe the problem and the work of Bernstein and Zelevinsky, while the proofs are outlined in [C2,C3], [He4], [W], as well as the introduction to [HT].

[AC] Arthur, J. and L. Clozel,, *Simple algebras, base change, and the advanced theory of the trace formula*, *Annals of Math. Studies*, **120**, Princeton: Princeton University Press (1989).

[B] Berkovich, V.G., Étale cohomology for non-archimedean analytic spaces, *Publ. Math. I.H.E.S.*, **78**, 5–161 (1993).

[2] A wildly ramified irreducible representation of W_F is one that remains irreducible upon restriction to P_F; a wildly ramified supercuspidal is one not isomorphic to its twist by any non-trivial unramified character.

[Bo] Boyer, P., Mauvaise réduction de variétés de Drinfeld et correspondance de Langlands locale, *Invent. Math.*, **138**, 573–629 (1999).

[BHK] Bushnell, C., G. Henniart, and P. Kutzko, Local Rankin-Selberg convolutions for GL_n: explicit conductor formula, *J. Amer. Math. Soc.*, **11**, 703–730 (1998).

[BK] Bushnell, C. and P. Kutzko, The admissible dual of $GL(N)$ via compact open subgroups, *Annals of Math. Studies*, **129** (1993).

[BH] Bushnell, C. and G. Henniart, Local tame lifting for $GL(n)$, I. *Publ. Math. IHES*, (1996); II: wildly ramified supercuspidals, *Astérisque*, **254** (1999).

[C1] Carayol, H., Non-abelian Lubin-Tate theory, in L. Clozel and J.S. Milne, eds., *Automorphic Forms, Shimura varieties, and L-functions*, New York: Academic Press, vol II, 15–39 (1990).

[C2] Carayol, H., Variétés de Drinfeld compactes, d'après Laumon, Rapoport, et Stuhler, Séminaire Bourbaki exp. 756 (1991-1992), *Astérisque* **206** (1992), 369–409.

[C3] Carayol, H., Preuve de la conjecture de Langlands locale pour GL_n: Travaux de Harris-Taylor et Henniart, Séminaire Bourbaki exp. 857. (1998-1999).

[Cl1] Clozel, L., Représentations Galoisiennes associées aux représentations automorphes autoduales de GL(n), *Publ. Math. I.H.E.S.*, **73**, 97–145 (1991).

[Cl2] Clozel, L., Automorphic forms and the distribution of points on odd-dimensional spheres, (manuscript, 2001).

[CL] Clozel, L. and J.-P. Labesse, Changement de base pour les représentations cohomologiques de certains groupes unitaires, appendix to J.-P. Labesse, *Cohomologie, stabilisation, et changement de base*, *Astérisque*, **257** (1999).

[D] Deligne, P., Les constantes des équations fonctionnelles des fonctions L, *Modular Functions of one variable II, Lect. Notes Math.*, **349**, 501–595 (1973).

[DKV] Deligne, P., D.Kazhdan, and M.-F.Vigneras, Représentations des algèbres centrales simples p-adiques, in J.-N.Bernstein, P.Deligne, D.Kazhdan, M.-F.Vigneras, *Représentations des groupes réductifs sur un corps local*, Paris: Hermann (1984).

[D1] Drinfeld, V., Elliptic modules, *Math. USSR Sbornik*, **23**, 561–592 (1974).

[D2] Drinfeld, V., Coverings of p-adic symmetric domains, *Fun. Anal. Appl.*, **10**, 107–115 (1976).

[F1] Faltings, G., The trace formula and Drinfeld's upper halfplane, *Duke Math. J.*, **76**, 467–481 (1994).

[F2] Faltings, G., A relation between two moduli spaces studied by V. G. Drinfeld, preprint, 2001.

[Fa] Fargues, L., Correspondances de Langlands locales dans la cohomologie des espaces de Rapoport-Zink, thèse de doctorat, Université Paris 7 (2001).

[Fu] Fujiwara, K., Rigid geometry, Lefschetz-Verdier trace formula and Deligne's conjecture, *Invent. Math.*, **127**, 489–533 (1997).

[H1] Harris, M., Supercuspidal representations in the cohomology of Drinfel'd upper half spaces; elaboration of Carayol's program, *Invent. Math.*, **129**, 75–119 (1997).

[H2] Harris, M., The local Langlands conjecture for $GL(n)$ over a p-adic field, $n < p$, *Invent. Math.*, **134**, 177–210 (1998).

[H3] Harris, M., Local Langlands correspondences and vanishing cycles on Shimura varieties, Proceedings of the European Congress of Mathematics, Barcelona, 2000. *Progress in Mathematics*, **201**, 407–427 (2001).

[HT] Harris, M. and R. Taylor, On the geometry and cohomology of some simple Shimura varieties, *Annals of Math. Studies*, **161** (2002).

[Hau] Hausberger, T. Représentations cuspidales dans la cohomologie des revêtements de Drinfeld: preuve de la conjecture de Drinfeld-Carayol en égale caractéristique, preprint (2001).

[He1] Henniart, G., La conjecture de Langlands locale numérique pour $GL(n)$, *Ann. scient. Ec. Norm. Sup*, **21**, 497–544 (1988).

[He2] Henniart, G., Caractérisation de la correspondence de Langlands locale par les facteurs ϵ de paires, *Invent. Math.*, **113**, 339–350 (1993).

[He3] Henniart, G., Une preuve simple des conjectures de Langlands pour $GL(n)$ sur un corps p-adique, *Invent. Math.*, (2000).

[He4] Henniart, G., A Report on the Proof of the Langlands Conjectures for $GL(N)$ over p-adic Fields, Current Developments in Mathematics 1999, International Press, 55–68 (1999).

[He5] Henniart, G., Sur la conjecture de Langlands locale pour GL_n, *J. Théorie des Nombres de Bordeaux, 13*, 167–187 (2001).

[Hu] Huber, R., Swan representations associated with rigid analytic curves, *J. Reine Angew. Math.*, **537**, 165–234 (2001).

[JPS] Jacquet, H., I. I. Piatetski-Shapiro, and J. Shalika, Rankin-Selberg convolutions, *Am. J. Math.*, **105**, 367–483 (1983).

[JS] Jiang, D. and D. Soudry, Generic representations and local Langlands reciprocity law for p-adic SO_{2n+1}, preprint (2001).

[K] Kottwitz, R., On the λ-adic representations associated to some simple Shimura varieties, *Invent. Math.*, **108**, 653–665 (1992).

[Ku] Kudla, S., The local Langlands correspondence: The non-archimedean case, *Proc. Symp. Pure Math.*, **55**, part 2, 365–391 (1994).

[LRS] Laumon, G., M. Rapoport, and U. Stuhler, $\mathcal{D}$-elliptic sheaves and the Langlands correspondence, *Invent. Math.*, **113**, 217–338 (1993).

[M] Mantovan, E., On certain unitary group Shimura varieties, Harvard Ph. D. thesis (2002).

[R] Rapoport, M., Non-archimedean period domains, Proceedings of the International Congress of Mathematicians, Zürich, 1994, 423–434. (1995).

[Rd] Rodier, F., Représentations de $GL(n, k)$ où k est un corps p-adique, *Astérisque*, **92-93** (1982), 201-218; Sém Bourbaki 1981-82, exposé no. 583.

[RZ] Rapoport, M. et T. Zink, *Period Spaces for p-divisible Groups*, Princeton: Annals of Mathematics Studies **141** (1996).

[Ro] Rogawski, J., Representations of $GL(n)$ and division algebras over a p-adic field, *Duke Math. J.*, **50**, 161–196 (1983).

[Sh] Shahidi, F., Local coefficients and normalization of intertwining operators for $GL(n)$, *Comp. Math.*, **48**, 271–295 (1983).

[S] Strauch, M., On the Jacquet-Langlands correspondence in the cohomology of the Lubin-Tate deformation tower, Preprintreihe SFB 478 (Münster), **72**, (1999).

[SS] Schneider, P. and U. Stuhler, The cohomology of p-adic symmetric spaces, *Invent. Math.*, **105**, 47–122 (1991).

[W] Wedhorn, T., The local Langlands correspondence for $GL(n)$ over p-adic fields, lecture at the summer school on Automorphic Forms on GL(n) at the ICTP Trieste, ICTP Lecture Notes Series (to appear).

[Z] Zelevinsky, A. V., Induced representations of reductive p-adic groups II: on irreducible representations of $GL(n)$, *Ann. Sci. E.N.S.*, **13**, 165–210 (1980).

ICM 2002 · Vol. II · 599–613

Vector Bundles, Linear Representations, and Spectral Problems

Alexander Klyachko*

Abstract

This paper is based on my talk at ICM on recent progress in a number of classical problems of linear algebra and representation theory, based on new approach, originated from geometry of stable bundles and geometric invariant theory.

2000 Mathematics Subject Classification: 14F05, 14M15, 14M17, 14M25, 15A42.
Keywords and Phrases: Bundles, Linear representations, Spectral problems.

1. Introduction

Theory of vector bundles brings a new meaning and adds a delicate geometric flavour to classical spectral problems of linear algebra, relating them to geometric invariant theory, representation theory, Schubert calculus, quantum cohomology, and various moduli spaces. The talk may be considered as a supplement to that of Hermann Weyl [35] from which I borrow the following quotation

"In preparing this lecture, the speaker has assumed that he is expected to talk on a subject in which he had some first-hand experience through his own work. And glancing back over the years he found that the one topic to which he has returned again and again is the problem of eigenvalues and eigenfunctions in its various ramifications."

2. Spectra and representations

Let's start with two classical and apparently independent problems.

Hermitian spectral problem. Find all possible spectra $\lambda(A+B)$ of sum of Hermitian operators A, B with given spectra

$$\lambda(A):\ \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A),$$
$$\lambda(B):\ \lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_n(B).$$

*Department of Mathematics, Bilkent University, Bilkent 06533, Ankara, Turkey. E-mail: klyachko@fen.bilkent.edu.tr

Among commonly known restrictions on spectra are *trace identity*

$$\sum_i \lambda_i(A+B) = \sum_j \lambda_j(A) + \sum_k \lambda_k(B)$$

and a number of classical inequalities, like that of Weyl [34]

$$\lambda_{i+j-1}(A+B) \le \lambda_i(A) + \lambda_j(B). \tag{2.0}$$

Tensor product problem. Find all components $V_\gamma \subset V_\alpha \otimes V_\beta$ of tensor product of two irreducible representations of GL_n with highest weights (=Young diagrams)

$$\alpha : a_1 \ge a_2 \ge \cdots \ge a_n$$
$$\beta : b_1 \ge b_2 \ge \cdots \ge b_n.$$

In contrast to the spectral problem (2.1) the coefficients of tensor product decomposition

$$V_\alpha \otimes V_\beta = \sum_\gamma c^\gamma_{\alpha\beta} V_\gamma \tag{2.1}$$

can be evaluated algorithmically by *Littlewood–Richardson rule*, which may be described as follows. Fill i-th row of diagram β by symbol i. Then $c^\gamma_{\alpha\beta}$ is equal to number of ways to produce diagram γ by adding cells from β to α in such a way that the symbols

i) weakly increase in rows,
ii) strictly increase in columns,
iii) reading all the symbols from right to left, and from top to bottom produces a *lattice permutation*, i.e. in every initial interval symbol i appears at least as many times as $i+1$.

It turns out that these two problems are essentially equivalent and have the same answer. To give it, let's associate with a subset $I \subset \{1, 2, \dots, n\}$ of cardinality $p = |I|$ Young diagram σ_I in a rectangular of format $p \times q$, $p+q=n$, cut out by polygonal line Γ_I, connecting SW and NE corners of the rectangular, with i-th unit edge running to the North, for $i \in I$, and to the East otherwise. One can formally multiply the diagrams by L-R rule

$$\sigma_I \sigma_J = \sum_k c^K_{IJ} \sigma_K \tag{2.2}$$

where $\sigma^K_{IJ} := c^{\sigma_K}_{\sigma_I \sigma_J}$ are L-R coefficients. Geometrically (2.2) is decomposition of product of two Schubert cycles in cohomology ring of Grassmannian G^q_p of linear subspaces of dimension p and codimension q.

Theorem 2.1. *The following conditions are equivalent*

i) *There exist Hermitian operators* A, B, $C = A + B$ *with spectra* $\lambda(A)$, $\lambda(B)$, $\lambda(C)$.

ii) *Inequality*

$$\lambda_K(C) \le \lambda_I(A) + \lambda_J(B), \tag{IJK}$$

holds each time L-R coefficient $c_{IJ}^K \neq 0$. *Here* $I, J, K \subset \{1, 2, \dots, n\}$ *are subsets of the same cardinality* $p < n$, *and* $\lambda_I(A) = \sum_{i \in I} \lambda_i(A)$.

iii) *For integer spectra* $\alpha = \lambda(A)$, $\beta = \lambda(B)$, $\gamma = \lambda(C)$ *the above conditions are equivalent to*

$$V_\gamma \subset V_\alpha \otimes V_\beta. \tag{2.3}$$

Remarks 2.2. (1) The last claim iii) implies a *recurrence procedure* to generate all α, β, γ with $c_{\alpha\beta}^\gamma \neq 0$:

$$c_{\alpha\beta}^\gamma \neq 0 \underset{\text{LR}}{\iff} V_\gamma \subset V_\alpha \otimes V_\beta \underset{\text{Th}}{\iff} \gamma_K \le \alpha_I + \beta_J \text{ each time } c_{IJ}^K \neq 0.$$

Here $c_{\alpha\beta}^\gamma$ are Littlewood-Richardson coefficients for group GL_n, while c_{IJ}^K are L-R coefficient for group GL_p of *smaller* rank $p < n$. An explicit form of this recurrence has been conjectured by A. Horn [13] in the framework of Hermitian spectral problem.

(2) Inequalities (IJK) for $c_{IJ}^K \neq 0$ define a cone in the space of triplets of spectra, and the facets of this cone correspond to $c_{IJ}^K = 1$. P. Belkale [3] was first to note that all inequalities (IJK) follows from those with $c_{IJ}^K = 1$, and in recent preprint A. Knutson, T. Tao, and Ch. Woodward [23] proved their independence. In my original paper [19] condition (2.3) appears in a weaker form

$$V_{N\gamma} \subset V_{N\alpha} \otimes V_{N\beta} \quad \text{for some } N > 0, \tag{2.3$'$}$$

and its equivalence to (2.3), known as *saturation conjecture*, was later proved by A. Knutson and T. Tao [22], and in more general quiver context by H. Derksen and J. Weyman [6].

Note that inequalities (IJK), although complete, are too numerous to be practical for large n. That is why L-R rule, in its different incarnations [22, 11], often provides a more intuitive way to see possible spectra for sum of Hermitian operators.

Example 2.3. Let A be Hermitian matrix with integer spectrum $\lambda(A) : \ a_1 \ge a_2 \ge \dots \ge a_n$ and $B \ge 0$ be a nonnegative matrix of rank one with spectrum $\lambda(B) : \ b \ge 0 \ge \dots \ge 0$. Viewing the spectra as Young diagrams, and applying L-R rule we find out that $\lambda(A) \otimes \lambda(B)$ is a sum of diagrams $\gamma : \ c_1 \ge c_2 \ge \dots \ge c_n$ satisfying the following *intrlacing inequalities*

$$c_1 \ge a_1 \ge c_2 \ge a_2 \cdots \ge c_n \ge a_n.$$

By Theorem 2.3 this implies *Cauchy interlacing theorem* for spectra

$$\lambda_i(A) \le \lambda_i(A+B) \le \lambda_{i-1}(A), \quad \mathrm{rk}B = 1, \quad B \ge 0,$$

known in mechanics as *Rayleigh-Courant-Fisher* principle: Let mechanical system S' is obtained from another one S, by imposing a linear constraint, e.g. by fixing a point of a drum. Then spectrum of S separates spectrum of S'.

3. Toric bundles

Historically Theorem 2.3 first appears as a byproduct of theory of toric vector bundles and sheaves, originated in [15, 17]. See other expositions of the theory in [21, 30], and further applications in [16, 33]. Vector bundles form a cross point at which the diverse subjects of this paper meet together.

3.1. Filtrations

To avoid technicalities let's consider the simplest case of projective plane

$$\mathbb{P}^2 = \{(x^\alpha : x^\beta : x^\gamma) | x \in \mathbb{C}\}$$

on which diagonal torus

$$T = \{(t_\alpha : t_\beta : t_\gamma) | t \in \mathbb{C}^*\} \tag{3.1}$$

acts by the formula

$$t \cdot x = (t_\alpha x^\alpha : t_\beta x^\beta : t_\gamma x^\gamma).$$

Orbits of this action are vertices, sides and complement of the coordinate triangle. In particular there is unique dense orbit, consisting of points with nonzero coordinates.

The objects of our interest are T-*equivariant* (or *toric* for short) vector bundles $\mathcal{E}$ over $\mathbb{P}^2$. This means that $\mathcal{E}$ is endowed with an action $T : \mathcal{E}$ which is linear on fibers and makes the following diagram commutative

$$\begin{array}{ccc} \mathcal{E} & \xrightarrow{t} & \mathcal{E} \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ \mathbb{P}^2 & \xrightarrow{t} & \mathbb{P}^2 \end{array} \qquad t \in T.$$

Let us fix a generic point $p_0 \in \mathbb{P}^2$ not in a coordinate line, and denote by

$$E := \mathcal{E}(p_0)$$

the corresponding generic fiber. There is no action of torus T on the fiber E. Instead the equivariant structure produces some distinguished subspaces in E by the following construction. Let us choose a generic point $p_\alpha \in X^\alpha$ in coordinate

line $X^\alpha : x^\alpha = 0$. Since T-orbit of p_0 is dense in $\mathbb{P}^2$, we can vary $t \in T$ so that tp_0 tends to p_α. Then for any vector $e \in E = \mathcal{E}(p_0)$, we have $te \in \mathcal{E}(tp_0)$ and can try the limit

$$\lim_{tp_0 \to p_\alpha} (te)$$

which either exists or not. Let us denote by $E^\alpha(0)$ the set of vectors $e \in E$ for which the limit exists:

$$E^\alpha(0) := \{e \in E | \lim_{tp_0 \to p_\alpha} (te) \quad \text{exists}\}.$$

Evidently $E^\alpha(0)$ is a vector subspace of E, independent of p_0 and p_α.

An easy modification of the previous construction allows to define for integer $m \in \mathbb{Z}$, the subspace

$$E^\alpha(m) := \left\{ e \in E | \lim_{tp_0 \to p_\alpha} \left(\frac{t_\alpha}{t_\beta}\right)^{-m} \cdot (te) \quad \text{exists} \right\}.$$

Roughly speaking $E^\alpha(m)$ consists of vectors $e \in E$ for which te vanishes up to order m as tp_0 tends to coordinate line X^α. The subspaces $E^\alpha(m)$ form a non-increasing exhaustive $\mathbb{Z}$-filtration:

$$\begin{aligned} E^\alpha : \cdots \supset E^\alpha(m-1) \supset E^\alpha(m) \supset E^\alpha(m+1) \supset \cdots, \\ E^\alpha(m) = 0, \text{ for } m \gg 0, \\ E^\alpha(m) = E, \text{ for } m \ll 0. \end{aligned} \tag{3.2}$$

Applying this construction to other coordinate lines, we get a triple of filtrations E^α, E^β, E^γ in generic fiber $E = \mathcal{E}(p_0)$, associated with toric bundle $\mathcal{E}$.

Theorem 3.1. *The correspondence*

$$\mathcal{E} \mapsto (E^\alpha, E^\beta, E^\gamma) \tag{3.3}$$

establishes an equivalence between category of toric vector bundles on $\mathbb{P}^2$ *and category of triply filtered vector spaces.*

We'll use notation $\mathcal{E}(E^\alpha, E^\beta, E^\gamma)$ for toric bundle corresponding to triplet of filtrations $E^\alpha, E^\beta, E^\gamma$.

3.2. Stability

The previous theorem tells that every property or invariant of a vector bundle has its counterpart on the level of filtrations. For application to spectral problems the notion of stability of a vector bundle $\mathcal{E}$ is crucial. Recall that $\mathcal{E} \to \mathbb{P}^2$ is said to be Mumford–Takemoto *stable* iff

$$\frac{c_1(\mathcal{F})}{\operatorname{rk}\mathcal{F}} < \frac{c_1(\mathcal{E})}{\operatorname{rk}\mathcal{E}} \tag{3.4}$$

for every proper subsheaf $\mathcal{F} \subset \mathcal{E}$, and *semistable* if weak inequalities hold. Here $c_1(\mathcal{E}) = \deg \det \mathcal{E}$ is the first Chern class. Donaldson theorem [7] brings a deep geometrical meaning to this seemingly artificial definition: Every stable bundle carries unique Hermit-Einstein metric (with Ricci curvature proportional to metric).

Theorem 3.2. *Toric bundle $\mathcal{E} = \mathcal{E}(E^\alpha, E^\beta, E^\gamma)$ is stable iff for every proper subspace $F \subset E$ the following inequality holds*

$$\frac{1}{\dim F} \sum_{\substack{\nu=\alpha,\beta,\gamma \\ i\in\mathbb{Z}}} i \dim F^{[\nu]}(i) < \frac{1}{\dim E} \sum_{\substack{\nu=\alpha,\beta,\gamma \\ i\in\mathbb{Z}}} i \dim E^{[\nu]}(i) \tag{3.5}$$

where $F^\nu(i) = F \cap E^\nu(i)$ is induces filtration with composition factors $F^{[\nu]}(i) = F^\nu(i)/F^\nu(i+1)$.

There is nothing surprising in this theorem since the sums in (3.5) are just Chern classes of the corresponding toric bundles and sheaves.

Remark 3.3. Inequality (3.5) depends only on *relative positions* of subspace $F \subset E$ with respect to filtrations $E^\alpha, E^\beta, E^\gamma$, which are given by three *Schubert cells* $s_\alpha, s_\beta, s_\gamma$. Hence we have one inequality each time.

$$s_\alpha \cap s_\beta \cap s_\gamma \neq \emptyset. \tag{3.6}$$

For filtrations in general position (3.6) is equivalent to nonvanishing of the product of *Schubert cycles* $\sigma_\alpha \cdot \sigma_\beta \cdot \sigma_\gamma \neq 0$ in cohomolgy ring of Grassmannian, and in this case stability inequalities (3.5) amount to inequalities (IJK) of Theorem 2.1.

3.3. Back to Hermitian operators

Let now E be Hermitian space and $H : E \to E$ be Hermitian operator with *spectral filtration*

$$E^H(x) = \begin{pmatrix} \text{sum of eigenspaces of } H \\ \text{with eigenvalues at least } x \end{pmatrix}. \tag{3.7}$$

The operator can be recovered from the filtration using *spectral decomposition*

$$H = \int_{-\infty}^{\infty} x dP_H(x)$$

where $P_H(x)$ is orthogonal projector with kernel $E^H(x)$. So in Hermitian space we have equivalence

$$\textit{Hermitian operators } = \mathbb{R}\textit{-filtrations.}$$

Let H^α be Hermitian operator with spectral filtration E^α. Its spectrum depends only on filtration E^α, and we define $\operatorname{Spec} E^\alpha := \operatorname{Spec} H^\alpha$.

Theorem 3.3. *Indecomposable triplet of $\mathbb{R}$-filtrations $E^\alpha, E^\beta, E^\gamma$ is stable iff there exists a Hermitian metric in E such that the sum of the corresponding Hermitian operators is a scalar*

$$H^\alpha + H^\beta + H^\gamma = \text{scalar}. \tag{3.8}$$

This is a toric version of Donaldson theorem on existence of Hermit–Einstein metric in stable bundles. Together with Theorem 3.2 it reduces solution of Hermitian spectral problem to stability inequalities (3.5), which by remark 3.3 amounts to inequalities (IJK) of Theorem 2.1.

See also Faltings talk [9] on arithmetical applications of stable filtrations.

3.4. Components of tensor product

In the previous section we explain that stability inequalities (3.5) ($\Leftrightarrow$ (IJK)) via toric Donaldson-Yau theorem solve Hermitian spectral problem. To relate this with tensor product part of Theorem 2.1 we need another interpretation of the stability inequalities via Geometric Invariant Theory [26].

Recall, that point $x \in \mathbb{P}(V)$ is said to be *GIT stable* with respect to linear action $G : V$ if G-orbit of the corresponding vector $\overline{x} \in V$ is closed and its stabilizer is finite. Let

$$X = \mathcal{F}^\alpha \times \mathcal{F}^\beta \times \mathcal{F}^\gamma$$

be product of three flag varieties of the same types as flags of the filtrations $E^\alpha, E^\beta, E^\gamma$, and $\mathcal{L}^\alpha$ be line bundle on the flag variety $\mathcal{F}^\alpha$ induced by character

$$\omega_\alpha : \mathrm{diag}(x_1, x_2, \dots, x_n) \mapsto x_1^{a_1} x_1^{a_2} \cdots x_1^{a_n},$$

where $\alpha : \ a_1 \geq a_2 \geq \cdots \geq a_n$ is the spectrum of filtration E^α, i.e. spectrum of the corresponding operator H^α.

Observation 3.4. *Vector bundle $\mathcal{E} = \mathcal{E}(E^\alpha, E^\beta, E^\gamma)$ is stable iff the corresponding triplet of flags*

$$x = F^\alpha \times F^\beta \times F^\gamma \in \mathcal{F}^\alpha \times \mathcal{F}^\beta \times \mathcal{F}^\gamma = X \hookrightarrow \mathbb{P}(\Gamma(X, \mathcal{L}))$$

is a GIT stable point w.r. to group $\mathrm{SL}(E)$ *and polarization* $\mathcal{L} = \mathcal{L}^\alpha \boxtimes \mathcal{L}^\beta \boxtimes \mathcal{L}^\gamma$.

This observation is essentially due to Mumford [25]. Notice that by Borel-Weil-Bott theorem [5] the space of global sections $\Gamma(\mathcal{F}^\alpha, \mathcal{L}^\alpha) = V_\alpha$ is just an irreducible representation of $\mathrm{SL}(E)$ with highest weight α. Hence $\Gamma(X, \mathcal{L})) = V_\alpha \otimes V_\beta \otimes V_\gamma$. Every stable vector $\overline{x}$ can be separated from zero by a G-invariant section of $\mathcal{L}^N$. Therefore triplet of flags in generic position is stable iff $[V_{N\alpha} \otimes V_{N\beta} \otimes V_{N\gamma}]^{\mathrm{SL}(E)} \neq 0$ for some $N \geq 1$. This proves the last part of Theorem 2.1, modulo the saturation conjecture.

4. Unitary operators and parabolic bundles

We have seen in the previous section that solution of the Hermitian spectral problem amounts to stability condition for toric bundles. A remarkable ramification of this idea was discovered by S. Angihotri and Ch. Woodward [2] for unitary spectral problem.

Let $U \in \mathrm{SU}(n)$ be unitary matrix with unitary spectrum

$$\varepsilon(U) = (e^{2\pi i \lambda_1}, e^{2\pi i \lambda_2}, \dots, e^{2\pi i \lambda_n}).$$

Let's normalize exponents λ_i as follows

$$\lambda(U) := \begin{cases} \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \\ \lambda_1 + \lambda_2 + \cdots + \lambda_n = 0, \\ \lambda_1 - \lambda_n < 1, \end{cases} \tag{4.1}$$

and, admitting an abuse of language, call $\lambda(U)$ *spectrum* of U.

Unitary spectral problem. Find possible spectra of product $\lambda(UV)$, when spectra of the factors $\lambda(U), \lambda(V)$ are given.

To state the result we need in *quantum cohomology* $\mathrm{H}^*_q(G^r_p)$ of Grassmannian G^r_p of linear subspaces of dimension p and codimension r. This is an algebra over polynomial ring $\mathbb{C}[q]$ generated by Schubert cycles σ_I, $I \subset \{1,2,\dots,n\}$, $|I| = p$, $n = p + r$ with multiplication given by the formula

$$\sigma_I * \sigma_J = \sum_{K,d} c^K_{IJ}(d) q^d \sigma_K$$

where structure constants $c^K_{IJ}(d)$ are defined as follows. Let $G^p_q \hookrightarrow \mathbb{P}(\bigwedge^p \mathbb{C}^n)$ be Plücker imbedding and

$$\varphi : \mathbb{P}^1 \to G^r_p$$

be a rational curve of degree d in Gressmanian $G^p_q \subset \mathbb{P}(\bigwedge^p \mathbb{C}^n)$. One can check that φ depends on $\dim G^p_q + nd$ parameters. For fixed point $x \in \mathbb{P}^1$ the condition $\varphi(x) \in \sigma_I$ imposes $\operatorname{codim} \sigma_I$ constraints on φ. Hence for

$$\operatorname{codim} \sigma_I + \operatorname{codim} \sigma_J + \operatorname{codim} \sigma_K = \dim G^r_p + nd$$

the numbers

$$(\sigma_I, \sigma_J, \sigma_K)_d = \#\{\varphi : \mathbb{P}^1 \to G^r_p \mid \varphi(x) \in \sigma_I, \varphi(x) \in \sigma_J, \varphi(x) \in \sigma_K, \deg \varphi = d\}$$

supposed to be finite. They are known as *Gromov - Witten invariants* and related to the structure constants by the formula

$$c^K_{IJ}(d) = (\sigma_I, \sigma_J, \sigma_{K^*})_d$$

where $K^* = \{n + 1 - k \mid k \in K\}$. For $d = 0$ they are just conventional Littlewood–Richardson coefficients c^K_{IJ}.

Theorem 4.1. *The following conditions are equivalent*

i) *There exist unitary matrices $W = UV$ with given spectra $\lambda(U), \lambda(V), \lambda(W)$.*

ii) *The inequality*

$$\lambda_I(U) + \lambda_J(V) \le d + \lambda_K(W) \qquad (\mathrm{IJK})_d$$

holds each time $c^K_{IJ}(d) \neq 0$.

4.1. Parabolic bundles

As in the Hermitian case solution of the unitary problem comes from its *holomorphic* interpretation in terms of vector bundles. To explain the idea let's start with vector bundle $\mathcal{E}$ over *compact* Riemann surface $\overline{X}$ of genus $g \ge 2$. It has unique topological invariant $c_1(\mathcal{E}) = \deg \det \mathcal{E}$, which for simplicity we suppose to be zero,

i.e. $\mathcal{E}$ be topologically trivial. Narasimhan-Seshadri theorem [27] claims that every stable bundle carries unique flat metric, and hence defines unitary monodromy representation

$$\rho_{\mathcal{E}} : \pi_1(\overline{X}, x_0) \to \mathrm{SU}(E), \quad E = \mathcal{E}(x_0).$$

This gives rise to equivalence

$$\mathcal{M}_g := \begin{pmatrix}\text{stable bundles}\\ \text{of degree zero}\end{pmatrix} = \begin{pmatrix}\text{irreducible uitary represen-}\\ \text{tations } \rho : \pi_1 \to \mathrm{SU}(E)\end{pmatrix}. \tag{4.2}$$

This theorem is an ancestor of the Donaldson-Yau generalization [7] to higher dimensions, and may be seen as a geometric version of Langlands correspondence.

In algebraic terms the theorem describes stable bundles in terms of solution of equation

$$[U_1, V_1][U_2, V_2] \cdots [U_g, V_g] = 1$$

in unitary matrices $U_i, V_j \in \mathrm{SU}(E)$. This is not the matrix problem we are currently interested in. To modify it let's consider *punctured* Riemann surface $X = \overline{X} \backslash \{p_1, p_2, \dots, p_\ell\}$. It has distinguished classes

$$\gamma_\alpha = (\text{small circle around } p_\alpha)$$

in fundamental group $\pi_1(X)$, and we can readily define an analogue of RHS of (4.2):

$$\mathcal{M}_g(\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(\ell)}) = \{\rho : \pi_1(X) \to \mathrm{SU}(E) \mid \lambda(\rho(\gamma_\alpha)) = \lambda^{(\alpha)}\}, \tag{4.3}$$

where $\lambda^{(\alpha)}$ is a given spectrum of monodromy around puncture p_α. C. S. Seshadri [31] manages to find an analogue of more subtle *holomorphic* LHS of (4.2) in terms of so called *parabolic bundles*.

Parabolic bundle $\mathcal{E}$ on X is actually a bundle on compactification $\overline{X}$ together with $\mathbb{R}$-filtration in every special fiber $E^\alpha = \mathcal{E}(p_\alpha)$ with support in an interval of length ≤ 1. The filtration is a substitution for spectral decomposition of $\rho(\gamma_\alpha)$, cf. (4.1). Seshadri also defines *(semi)stability* of parabolic bundle $\mathcal{E}$ by inequalities

$$\frac{\mathrm{Par}\deg \mathcal{F}}{\mathrm{rk}\,\mathcal{F}} \leq \frac{\mathrm{Par}\deg \mathcal{E}}{\mathrm{rk}\,\mathcal{E}}, \quad \forall \mathcal{F} \subset \mathcal{E}, \tag{4.4}$$

where the parabolic degree is given by equation $\mathrm{Par}\deg \mathcal{E} = \deg \mathcal{E} + \sum_{\alpha,i} \lambda_i^{(\alpha)}$. Metha-Seshadri theorem [24] claims that every stable parabolic bundle $\mathcal{E}$ on X carries unique flat metric with given spectra of monodromies $\lambda(\gamma_\alpha) = \lambda^{(\alpha)}$. This gives a holomorphic interpretation of the space (4.3)

$$\mathcal{M}_g(\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(\ell)}) = \begin{pmatrix}\text{stable parabolic bundles of degree zero}\\ \text{with given types of the filtrations}\end{pmatrix}. \tag{4.5}$$

In the simplest case of projective line with three punctures (4.3) amounts to space of solutions of equation $UVW = 1$ in unitary matrices $U, V, W \in \mathrm{SU}(n)$ with given

spectra. By Metha-Seshadry theorem solvability of this equation is equivalent to stability inequalities (4.4). In the case under consideration holomorphic vector bundle $\mathcal{E}$ on $\mathbb{P}^1$ is trivial, $\mathcal{E} = E \times \mathbb{P}^1$, and hence its subbundle $\mathcal{F} \subset \mathcal{E}$ of rank p is nothing but a rational curve $\varphi : \mathbb{P}^1 \to G_p(E)$ in Grassmannian. This allows to write down stability condition (4.4) in terms of quantum cohomology, and eventually arrive at Theorem 4.1.

5. Further ramifications

The progress in Hermitian and unitary spectral problems open way for solution of a variety of others classical, and not so classical, problems. Most of them, however, have no holomorphic interpretation, and require different methods, borrowed from harmonic analysis on homogeneous spaces, symplectic geometry, and geometric invariant theory.

5.1. Multiplicative singular value problem

The problem in question is about possible singular spectrum $\sigma(AB)$ of product of complex matrices with given singular spectra $\sigma(A)$ and $\sigma(B)$. Recall, that singular spectrum of complex matrix A is spectrum of its radial part $\sigma(A) := \lambda(\sqrt{A^*A})$.

For a long time it was observed that every inequality for Hermitian problem has a *multiplicative* counterpart for the singular one. For example multiplicative version of Weyl's inequality $\lambda_{i+j-1}(A+B) \leq \lambda_i(A) + \lambda_j(B)$ is $\sigma_{i+j-1}(AB) \leq \sigma_i(A)\sigma_j(B)$. The equivalence between these two problems was conjectured by R. C. Thompson, and first proved by the author [20] using harmonic analysis on symmetric spaces. Later on A. Alekseev, E. Menreken, and Ch. Woodward [1] gave an elegant conceptual solution based on Drinfeld's Poisson-Lie groups [8]. Here is a precise statement for classical groups.

Theorem 5.1. *Let $\mathbb{G}$ be one of the classical groups $\mathrm{SL}(n,\mathbb{C})$, $\mathrm{SO}(n,\mathbb{C})$, or $\mathrm{Sp}(2n,\mathbb{C})$ and $\mathbb{L}$ be the corresponding compact Lie algebra of traceless skew Hermitian complex, real, or quaternionic $n \times n$ matrices respectively. Then the following conditions are equivalent*

(1) *There exist $A_i \in \mathbb{G}$ with given singular spectra $\sigma(A_i) = \sigma_i$ and*

$$A_1 A_2 \cdots A_N = 1.$$

(2) *There exist $H_i \in \mathbb{L}$ with spectra $\lambda(H_i) = \sqrt{-1}\log\sigma_i$ and*

$$H_1 + H_2 + \cdots + H_N = 0.$$

Note, however, that neither of the above approaches solve the singular problem per se, but reduces it to Hermitian one. Both of them suggest that all three problems

must be treated in one package. More precisely, every compact simply connected group G give birth to three symmetric spaces

- The group G itself,
- Its Lie algebra L_G,
- The dual symmetric space $H_G = G^{\mathbb{C}}/G$,

of positive, zero, and negative curvature, and to three "spectral problems" concerned with support of convolution of G orbits in these spaces, see [20] for details. For $G = \mathrm{SU}(n)$ we return to the package of unitary, Hermitian, and singular problems.

The first two problems may be effectively treated in framework of vector bundles with structure group G, as explained in sections 2–4. Many flat, i.e. additive "spectral problem" has been solved by A. Berenstein and R. Sjammar in a very general setting [4].

5.2. Other symmetric spaces

As an example of unresolved problem let's consider symmetric spaces associated with different incarnations of Grassmannian

- Compact $U(p+q)/U(p) \times U(q)$,
- Flat $\mathrm{Mat}(p,q)$ = coomplex $p \times q$ matrices,
- Hyperbolic $U(p,q)/U(p) \times U(q)$.

In compact case the corresponding spectral problem is about possible angles between three p-subspaces $U, V, W \subset \mathbb{H}^n$ in Hermitian space $\mathbb{H}^n$ of dimension $n = p + q$, $p \leq q$. The *Jordan angles*

$$\widehat{UV} = (\varphi_1, \varphi_2, \dots, \varphi_p), \qquad 0 \leq \varphi \leq \frac{\pi}{2}$$

between subspaces U, V are defined via spectrum of product of orthogonal projectors $\pi_{UV} : U \to V$ and $\pi_{VU} : V \to U$

$$\lambda(\sqrt{\pi_{UV}\pi_{VU}}) : \cos\varphi_1 \geq \cos\varphi_2 \geq \cdots \geq \cos\varphi_p \geq 0.$$

Yu. Neretin [28] proved Lidskii type inequalities[1] for angles $\widehat{UV}, \widehat{VW}, \widehat{WU}$, and conjectured that other inequalities are the same as in the Hermitian case. Note, however, that the unitary triplet suggests existence of nonhomogeneous "quantum" inequalities, e.g. sum of angles of a spherical triangle is $\leq \pi$.

In flat case the problem is about relation between singular spectra of $p \times q$ matrices $\sigma(A-B)$, $\sigma(B-C)$, $\sigma(C-A)$. This *additive* singular problem was resolved by O'Shea and Sjamaar [29].

In hyperbolic case the question is about angles between maximal *positive* subspaces $U, V, W \subset \mathbb{H}^{pq}$ in Hermitian space of signature (p,q). They are defined by equation

$$\lambda(\sqrt{\pi_{UV}\pi_{VU}}) : \cosh\varphi_1 \geq \cosh\varphi_2 \geq \cdots \geq \cosh\varphi_p \geq 1.$$

[1] He actually deals with *real* Grassamnnian.

Again our experience with the unitary triplet suggests that the exponential map establishes a Thompson's type correspondence between O'Shea-Sjamaar inequalities for additive singular problem and that of for hyperbolic angles.

5.3. P-adic spectral problems

There is also a nonarchimedian counterpart of this theory, which deals with classical Chevalley groups $\mathbb{G}_p = \mathrm{SL}(n,\mathbb{Q}_p)$, $\mathrm{SO}(n,\mathbb{Q}_p)$, or $\mathrm{Sp}(2n,\mathbb{Q}_p)$ over p-adic field $\mathbb{Q}_p$ and their maximal compact subgroups $\mathbb{K}_p = \mathrm{SL}(n,\mathbb{Z}_p)$, $\mathrm{SO}(n,\mathbb{Z}_p)$, or $\mathrm{Sp}(2n,\mathbb{Z}_p)$ respectively. Double coset $\mathbb{K}_p g \mathbb{K}_p$ may be treated as a complete invariant of lattice $L = gL_0$, $L_0 = \mathbb{Z}_p^{\oplus n}$ with respect to $\mathbb{K}_p$. We call lattice $L = gL_0$ *unimodular*, *orthogonal* or *symplectic* if respectively $g \in \mathrm{SL}(n,\mathbb{Q}_p)$, $g \in \mathrm{SO}(n,\mathbb{Q}_p)$ or $g \in \mathrm{Sp}(2n,\mathbb{Q}_p)$.

It is commonly known that in the unimodular case there exists a basis e_i of L_0 such that $\tilde{e}_i = p^{a_i} e_i$ form a basis of L for some $a_i \in \mathbb{Z}$. We define *index* $(L : L_0)$ by

$$(L : L_0) = (p^{a_1}, p^{a_2}, \dots, p^{a_n}), \qquad a_1 \geq a_2 \geq \dots \geq a_n. \tag{5.1}$$

Notice that unimodularity $g \in \mathrm{SL}(n,\mathbb{Q}_p)$ implies $a_1 + a_2 + \dots + a_n = 0$.

The index $(L : L_0)$ of an orthogonal or a symplectic lattices has extra symmetries. In orthogonal case we may choose the above basis e_i of L_0 to be *neutral*, in which case the quadratic form becomes

$$\sum_{1-n}^{n-1} x_i x_{-i}, \quad i \equiv n-1 \mod 2.$$

Then the index takes the form

$$(L : L_0) = (p^{a_{n-1}}, p^{a_{n-3}}, \dots, p^{a_{3-n}}, p^{a_{1-n}}), \tag{5.2}$$

where $a_{n-1} \geq a_{n-3} \geq \dots \geq a_{3-n} \geq a_{1-n}$, and $a_{-i} = -a_i$.

Similarly, for *symplectic* lattice L we can choose symplectic basis e_i, f_j of L_0 such that $\tilde{e}_i = p^{a_i} e_i$ and $\tilde{f}_j = p^{-a_j} f_j$ form a basis of L. In this case we have

$$(L : L_0) = (p^{a_n}, p^{a_{n-1}}, \dots, p^{a_1}, p^{-a_1}, \dots, p^{-a_{n-1}}, p^{-a_n}), \tag{5.3}$$

with $a_n \geq a_{n-1} \geq, \dots, \geq a_1 \geq 0$.

Notice that the spectra (5.1)-(5.3) have the same symmetry, as singular spectrum $\sigma(A)$ of a matrix $A \in \mathbb{G}$ in the corresponding classical *complex* group.

Theorem 5.2. *The following conditions are equivalent*

(1) *There exists a sequence of (unimodular, orthogonal, symplectic) lattices*

$$L_0, L_1, \dots, L_{N-1}, L_N = L_0$$

of given indices $\sigma_i = (L_i : L_{i-1})$.

(2) *The indices* σ_i *satisfy the equivalent conditions of Theorem 5.1 for the corresponding complex group* $\mathbb{G}$.

We'll give proof elsewhere. The theorem is known for the unimodular lattices, see [10].

5.4. Final remarks

In the talk I try to trace the flaw of ideas from the theory of vector bundles to spectral problems. It seems C. Simpson [32] was the first to note that vector bundles technic has nontrivial implications in linear algebra. He proved that product $C_1C_2\cdots C_N$ of conjugacy classes $C_i \subset \mathrm{SL}(n,\mathbb{C})$ is dense in $\mathrm{SL}(n,\mathbb{C})$ iff

$$\begin{aligned} \dim C_1 + \dim C_2 + \cdots + \dim C_N &\geq (n+1)(n-2), \\ r_1 + r_2 + \cdots + r_N &\geq n, \end{aligned} \tag{5.4}$$

where r_i is maximal codimension of root space of a matrix $A_i \in C_i$. This problem was suggested by P. Deligne, who noted that under condition

$$\dim C_1 + \dim C_2 + \cdots + \dim C_N = 2n^2 - 2$$

an irreducible solution of equation $A_1A_2\cdots A_N = 1$, $a_i \in C_i$ is unique up to conjugacy, see book of N. Katz [14] on this rigidity phenomenon.

I think that inverse applications to moduli spaces of vector bundles are sill ahead. One may consider *polygon spaces* [18, 12] as a toy example of this feedback, corresponding to toric 2-bundles. A similar space of *spherical polygons* in $\mathbb{S}^3$ with given sides is a model for moduli space of flat connections in punctured Riemann sphere. Its description is a challenge problem.

There are many interesting results, e.g. infinite dimensional spectral problems, which fall out of this survey. I refer to Fulton's paper [10] for missing details.

References

[1] A. Alekseev, E. Meinrenken, & C. Woodward, Linearization of Poisson actions and singular values of matrix product, *Ann. Inst. Fourier (Grenoble),* 51 (2001), no. 6, 1691–1717.

[2] S. Angihotri & C. Woodward, Eigenvalues of products of unitary matrices and quantum Schubert calculus, *Math. Res. Letters,* 5 (1998), 817–836.

[3] P. Belkale, Local systems on $\mathbb{P}^1 - S$ for S a finite set, *Compositio Math.,* 129 (2001), no. 1, 67–86.

[4] A. Berenstein & R. Sjamaar, Coadjoint orbits, moment polytopes, and the Hilbert–Mumford criterion, *J. Amer. Math. Soc.,* 13 (2000), no. 2, 433–466.

[5] R. Bott, Homogeneous vector bundles, *Ann. of Math.,* 66 (1957), 203–248.

[6] H. Derksen & J. Weyman, Semi-invariants of quivers and saturation for Littlewood-Richardson theorem, *J. Amer. Math. Soc.,* 13 (2000), no. 3, 467–479.

[7] S. K. Donaldson, Infinite determinants, stable bundles and curvature, *Duke Math. J.,* 54 (1987), 231–247.

[8] V. G. Drinfeld, Quantum groups, *Proceedings of the International Congress of Mathematicians*, vol. 1,2 (Berkeley, 1986), Amer. Math. Soc., Providence, RI, 1987, 798–820.

[9] G. Faltings, Mumford-Stabilität in der algebraischen Geometrie, *Proceedings of the International Congress of Mathematiciens*, vol. 1,2, (Zürich, 1994), Birkhäuser, Basel, 1995, 648–655.

[10] W. Fulton, Eigenvalues, invariant factors, highest weights, and Schubert calculus, *Bull. Amer. Math. Soc.,* 37 (2000), no. 3, 209–249.

[11] O. Gleizer & A. Postnikov, Littlewood-Richardson coefficients via Yang-Baxter equation, *Internat. Math. Res. Notices* (2000), no. 14, 741–774.

[12] J.-C. Hausmann & A. Knutson, The cohomology ring of polygon spaces, *Ann. Inst. Fourier (Grenoble),* 48 (1998), no. 1, 281–321.

[13] A. Horn, Eigenvalues of sum of Hemitian matrices, *Pacific J. Math.,* 12 (1962), 225–241.

[14] N. M. Katz, *Rigid local systems*, Princeton University Press, Princeton, 1996.

[15] A. A. Klyachko, Equivariant bundles on toric varieties, *Izv. Akad. Nauk SSSR Ser. Mat.,* 53 (1989), no. 5, 1001–1039 (Russian); *Math. USSR-Izv.,* 35 (1990), no. 2, 63–64.

[16] A. A. Klyachko, Moduli of vector bundles and class numbers, *Functional. i Priozhen.,* 25 (1991), 81–83 (Russian); *Funct. Anal. Appl.,* 25 (1991), no. 1, 67–69.

[17] A. A. Klyachko, Vector bundles and torsion free sheaves on the projective plane, *Preprint Max-Planck-Institute fur Mathematik* MPI/91-59, (1991).

[18] A. A. Klyachko, Spatial polygons and stable configurations of points in the projective line, *Algebraic geometry and its applications (Yaroslavl, 1992)*, Vieweg, Braunschweig, 1994, 67–84.

[19] A. A. Kyachko, Stable bundles, repesentation theory and Hermitian operators, *Selecta Mathematica,* 4 (1998), 419–445.

[20] A. A. Klyachko, Random walks on symmetric spaces and and inequalities for matrix spectra, *Linear Algebra Appl.,* 319 (2000), no. 2-3, 37–59.

[21] A. Knutson & E. Sharp, Sheaves on toric varieties for physics, *Adv. Theor. Math. Phys.,* 2 (1998), no. 4, 873–961.

[22] A. Knutson & T. Tao, The honeycomb model of $\mathrm{GL}(n,\mathbb{C})$ tensor products. I. Proof of the saturation conjecture, *J. Amer. Math. Soc.,* 12 (1999), no. 2, 1055–1090.

[23] A. Knutson, T. Tao & Ch. Woodward, The honecomb model for $\mathrm{GL}(n,\mathbb{C})$ tensor products II: Facets of Littlewood–Richardson cone, *Preprint* (2001).

[24] V. B. Metha & C. S. Seshadri, Moduli of vector bundles on curves with parabolic sructure, *Math. Ann.,* 258 (1980), 205-239.

[25] D. Mumford, Projective invariants of projective structures, Proc. Int. Congress of Math. Sockholm, 1963, Almquist & Wiksells, Uppsala, 1963, 526–530.

[26] D. Mumford, J. Fogarty, & F. Kirwan, *Geometric invariant theory*, Springer, Berlin, 1994.

[27] M. S. Narasimhan & C. S. Seshadri, Stable and unitary vector bundles on a compact Riemann surface, *Ann. Math.,* 82 (1965), 540–567.

[28] Yu. Neretin, On Jordan angles and triangle inequality in Grassmannian, *Geom. Dedicata,* 86 (2001), no. 1-3, 81–92.

[29] L. O'Shea & R. Sjamaar, Moments maps and Riemannian symmetric pairs, *Math. Ann.,* 317 (2000), no. 3, 415–457.

[30] M. Perling, Graded rings and equivariant sheaves on toric varieties, *Preprint Univ. Kaiserslauten,* (2001).

[31] C. S. Seshadri, Moduli of vectir bundles on curves with parabolic structures, *Bull. Amer. Math. Soc.,* (1977), 124–126.

[32] C. T. Sympson, Product of Matrices, *Differential geometry, global analysis, and topology,* Canadian Math. Soc. Conf. Proc., vol. 12, AMS, Providence RI, 1992, 157–185.

[33] C. Vafa & E. Witten, *A strong coupling test of S-duality, Nuclear Phys. B,* 431 (1994), no. 1-2, 3–77.

[34] H. Weyl, Das asymptotischer Verteilungsgesetz der Eigenwerte lineare partialler Differentialgleichungen, *Math. Ann.,* 71 (1912), 441–479.

[35] H. Weyl, Ramifications, old and new, of the eigenvalue problem, *Bull. Amer. Math. Soc.,* 56 (1950), 115–139.

ICM 2002 · Vol. II · 615–627

Branching Problems of Unitary Representations

Toshiyuki Kobayashi*

Abstract

The irreducible decomposition of a unitary representation often contains continuous spectrum when restricted to a non-compact subgroup. The author singles out a nice class of branching problems where each irreducible summand occurs discretely with finite multiplicity (admissible restrictions). Basic theory and new perspectives of admissible restrictions are presented from both analytic and algebraic view points. We also discuss some applications of admissible restrictions to modular varieties and L^p-harmonic analysis.

2000 Mathematics Subject Classification: 22E46, 43A85, 11F67, 53C50, 53D20.
Keywords and Phrases: Unitary representation, Branching law, Reductive Lie group.

1. Introduction

Let π be an irreducible unitary representation of a group G. A **branching law** is the irreducible decomposition of π when restricted to a subgroup G':

$$\pi|_{G'} \simeq \int_{\widehat{G'}}^{\oplus} m_\pi(\tau)\tau \, d\mu(\tau) \qquad \text{(a direct integral)}. \tag{1.1}$$

Such a decomposition is unique, for example, if G' is a reductive Lie group, and the **multiplicity** $m_\pi : \widehat{G'} \to \mathbb{N} \cup \{\infty\}$ makes sense as a measurable function on the unitary dual $\widehat{G'}$.

Special cases of **branching problems** include (or reduce to) the followings: Clebsch-Gordan coefficients, Littlewood-Richardson rules, decomposition of tensor product representations, character formulas, Blattner formulas, Plancherel theorems for homogeneous spaces, description of breaking symmetries in quantum mechanics, theta-lifting in automorphic forms, etc. The restriction of unitary representations serves also as a method to study discontinuous groups for non-Riemannian homogeneous spaces (e.g. [Mg, Oh]).

*RIMS, Kyoto University, Kyoto 606-8502, Japan. E-mail: toshi@kurims.kyoto-u.ac.jp

Our interest is in the branching problems for (non-compact) reductive Lie groups $G \supset G'$. In this generality, there is no known algorithm to find branching laws. Even worse, branching laws usually contain both discrete and continuous spectrum with possibly infinite multiplicities (the multiplicity is infinite, for example, in the decomposition of the tensor product of two principal series representations of $SL(n,\mathbb{C})$ for $n \geq 3$, [Ge-Gr]).

The author introduced the notion of **admissible restrictions** and **infinitesimal discrete decomposability** in [Ko$_5$] and [Ko$_9$], respectively, seeking for a good framework of branching problems, in which we could expect especially a simple and detailed study of branching laws, which in turn might become powerful methods in other fields as well where restrictions of representations naturally arise.

The criterion in Theorem B indicates that there is a fairly rich examples of admissible restrictions; some are known and the others are new. In this framework, a number of explicit branching laws have been newly found (e.g. [D-Vs, Gr-W$_{1,2}$, Hu-P-S, Ko$_{1,3,4,8}$, Ko-Ø$_{1,2}$, Li$_2$, Lo$_{1,2}$, X]). The point here is that branching problems become accessible by algebraic techniques if there is no continuous spectrum.

The first half of this article surveys briefly a general theory of admissible restrictions both from analytic and algebraic view points (§2, §3). For the simplicity of exposition, we restrict ourselves to unitary representations, although a part of the theory can be generalized to non-unitary representations. The second half discusses some applications of discretely decomposable restrictions. The topics range from representation theory itself (§4) to some other fields such as L^p-analysis on non-symmetric homogeneous spaces (§5) and topology of modular varieties (§6).

2. Admissible restrictions to subgroups

Let G' be a subgroup of G, and $\pi \in \widehat{G}$. In light of (1.1), we introduce:

Definition 2.1. *We say the restriction $\pi|_{G'}$ is* **G'-admissible** *if it decomposes* **discretely** *and the multiplicity $m_\pi(\tau)$ is* **finite** *for any $\tau \in \widehat{G'}$.*

One can easily prove the following assertion:

Theorem A ([Ko$_5$, Theorem 1.2]). *Let $G \supset G' \supset G''$ be a chain of groups, and $\pi \in \widehat{G}$. If the restriction $\pi|_{G''}$ is G''-admissible, then $\pi|_{G'}$ is G'-admissible.*

Throughout this article, we shall treat the setting as below:

Definition 2.2. *We say (G,G') is a* **pair of reductive Lie groups** *if*

1) G is a real reductive linear Lie group or its finite cover, and

2) G' is a closed subgroup, and is reductive in G.

Then, we shall fix maximal compact subgroups $K \supset K'$ of $G \supset G'$, respectively.

A typical example is a **reductive symmetric pai** (G,G'), by which we mean that G is as above and that G' is an open subgroup of the set G^σ of the fixed points of an involutive automorphism σ of G. For example, $(G,G') = (GL(n,\mathbb{C}), GL(n,\mathbb{R}))$, $(SL(n,\mathbb{R}), SO(p,n-p))$ are the cases.

Let (G, G') be a pair of reductive Lie groups. Here are previously known examples of **admissible restrictions**:

Example 2.3. The restriction $\pi|_{G'}$ is G'-admissible in the following cases:

1) (Harish-Chandra's admissibility theorem) $\pi \in \widehat{G}$ is arbitrary and $G' = K$.
2) (Howe, [Ho$_1$]) π is the Segal-Shale-Weil representation of the metaplectic group G, and its subgroup $G' = G'_1 G'_2$ forms a dual pair with G'_1 compact.

In these examples, either the subgroup G' or the representation π is very special, namely, G' is compact or π has a highest weight. Surprisingly, without such assumptions, it can happen that the restriction $\pi|_{G'}$ is G'-admissible. The following criterion asserts that the "balance" of G' and π is crucial to the G'-admissibility.

Theorem B (criterion for admissible restrictions, [Ko$_7$]). *Let $G \supset G'$ be a pair of reductive Lie groups, and $\pi \in \widehat{G}$. If*

$$\mathrm{Cone}(G') \cap \mathrm{AS}_K(\pi) = \{0\}, \tag{2.1}$$

then the restriction $\pi|_{K'}$ is K'-admissible. In particular, the restriction $\pi|_{G'}$ is G'-admissible, namely, decomposes discretely with finite multiplicity.

A main tool of the proof of Theorem B is the microlocal study of characters by using the singularity spectrum of hyperfunctions. The idea goes back to Atiyah, Howe, Kashiwara and Vergne [A, Ho$_2$, Ks-Vr] in the late '70s. The novelty of Theorem B is to establish a framework of **admissible restrictions** with a number of new examples of interest, which rely on a deeper understanding of the unitary dual developed largely in the '80s (see [Kn-Vo] and references therein).

Let us briefly explain the notation used in Theorem B. We write $\mathfrak{k}'_0 \subset \mathfrak{k}_0$ for the Lie algebras of $K' \subset K$, respectively. Take a Cartan subalgebra $\mathfrak{t}_0$ of $\mathfrak{k}_0$. Then, $\mathrm{AS}_K(\pi)$ is the **asymptotic K-support** of π ([Ks-Vr]), and $\mathrm{Cone}(G')$ is defined as

$$\mathrm{Cone}(G') := \sqrt{-1}(\mathfrak{t}_0^* \cap \mathrm{Ad}^*(K)(\mathfrak{k}_0'^{\perp})). \tag{2.2}$$

By definition, both $\mathrm{AS}_K(\pi)$ and $\mathrm{Cone}(G')$ are closed cones in $\sqrt{-1}\mathfrak{t}_0^*$.

Example 2.4. If $G' = K$, then the assumption (2.1) is automatically fulfilled because $\mathrm{Cone}(G') = \{0\}$. The conclusion of Theorem B in this special case is nothing but **Harish-Chandra's admissibility theorem** (Example 2.3 (1)).

To apply Theorem B for non-compact G', we rewrite the assumption (2.1) more explicitly in specific settings. On the part $\mathrm{Cone}(G')$, we mention:

Example 2.5. $\mathrm{Cone}(G')$ is a linear subspace $\sqrt{-1}(\mathfrak{t}_0^*)^{-\sigma}$ (modulo the Weyl group) if (G, G') is a reductive symmetric pair given by an involution σ. Here, we have chosen a Cartan subalgebra $\mathfrak{t}_0$ to be maximally σ-split.

On the part $\mathrm{AS}_K(\pi)$, let us consider a unitary representation π_λ which is "attached to" an elliptic coadjoint orbit $\mathcal{O}_\lambda := \mathrm{Ad}^*(G)\lambda$, in the orbit philosophy due

to Kirillov-Kostant. This representation is a unitarization of a Zuckerman-Vogan module $A_{\mathfrak{q}}(\lambda)$ after some ρ-shift, and can be realized in the Dolbeault cohomology group on $\mathcal{O}_\lambda$ by the results of Schmid and Wong. (Here, we adopt the same polarization and normalization as in a survey [Ko$_4$, §2], for the **geometric quantization** $\mathcal{O}_\lambda \Rightarrow \pi_\lambda$.) We note that $\pi_\lambda \in \widehat{G}$ for "most" λ. Let $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ be the complexification of a Cartan decomposition of the Lie algebra $\mathfrak{g}_0$ of G. We set

$$\Delta_\lambda^+(\mathfrak{p}) := \{\alpha \in \Delta(\mathfrak{p}, \mathfrak{t}) : \langle \lambda, \alpha \rangle > 0\}, \quad \text{for } \lambda \in \sqrt{-1}\mathfrak{t}_0^*.$$

The original proof (see [Ko$_5$]) of the next theorem was based on an algebraic method without using microlocal analysis. Theorem B gives a simple and alternative proof.

Theorem C ([Ko$_5$]). *Let $\pi_\lambda \in \widehat{G}$ be attached to an elliptic coadjoint orbit $\mathcal{O}_\lambda$. If*

$$\mathbb{R}\text{-span } \Delta_\lambda^+(\mathfrak{p}) \cap \mathrm{Cone}(G') = \{0\}, \tag{2.2}$$

then the restriction $\pi_\lambda|_{G'}$ is G'-admissible.

Let us illustrate Theorem C in Examples 2.6 and 2.7 for non-compact G'. For this, we note that a maximal compact subgroup K is sometimes of the form $K_1 \times K_2$ (locally). This is the case if G/K is a Hermitian symmetric space (e.g. $G = Sp(n, \mathbb{R}), SO^*(2n), SU(p, q)$). It is also the case if $G = O(p, q)$, $Sp(p, q)$, etc.

Example 2.6 ($K \simeq K_1 \times K_2$)**.** Suppose K is (locally) isomorphic to the direct product group $K_1 \times K_2$. Then, *the restriction $\pi_\lambda|_{G'}$ is G'-admissible if $\lambda|_{\mathfrak{t} \cap \mathfrak{k}_2} = 0$ and $G' \supset K_1$*. So does the restriction $\pi|_{G'}$ if π is any subquotient of a coherent continuation of π_λ. This case was a prototype of G'-admissible restrictions $\pi|_{G'}$ (where G' is non-compact and π is a non-highest weight module) proved in 1989 by the author ([Ko$_1$; Ko$_2$, Proposition 4.1.3]), and was later generalized to Theorems B and C. Special cases include:

(1) $K_1 \simeq \mathbb{T}$, then π is a unitary highest weight module. The admissibility of the restrictions $\pi|_{G'}$ in this case had been already known in '70s (see Martens [Mt], Jakobsen-Vergne [J-Vr]).

(2) $K_1 \simeq SU(2)$, then π_λ is a quaternionic discrete series. Admissible restrictions $\pi|_{G'}$ in this case are especially studied by Gross and Wallach [Gr-W$_1$] in '90s.

(3) $K_1 \simeq O(q), U(q), Sp(q)$. Explicit branching laws of the restriction $\pi_\lambda|_{G'}$ for singular λ are given in [Ko$_3$, Part I] with respect to the vertical inclusions of the diagram below (see also [Ko$_1$, Ko$_5$] for those to horizontal inclusions).

$$\begin{array}{ccccc}
O(4p, 4q) & \supset & U(2p, 2q) & \supset & Sp(p, q) \\
\cup & & \cup & & \cup \\
O(4r) \times O(4p-4r, 4q) & \supset & U(2r) \times U(2p-2r, 2q) & \supset & Sp(r) \times Sp(p-r, q)
\end{array}$$

Example 2.7 (conformal group). There are 18 series of irreducible unitary representations of $G := U(2,2)$ with regular integral infinitesimal characters. Among them, 12 series (about "67% " !) are G'-admissible when restricted to $G' := Sp(1,1)$.

The assumption in Theorem B is in fact necessary. By using the technique of symplectic geometry, the author proved the converse statement of Theorem B:

Theorem D ([Ko$_{13}$]). *Let $G \supset G'$ be a pair of reductive Lie groups, and $\pi \in \widehat{G}$. If the restriction $\pi|_{K'}$ is K'-admissible, then* $\mathrm{Cone}(G') \cap \mathrm{AS}_K(\pi) = \{0\}$.

3. Infinitesimal discrete decomposability

The definition of admissible restrictions (Definition 2.1) is "analytic", namely, based on the direct integral decomposition (1.1) of unitary representations. Next, we consider discrete decomposable restrictions by a purely algebraic approach.

Definition 3.1 ([Ko$_9$, Definition 1.1]). *Let $\mathfrak{g}$ be a Lie algebra. We say a $\mathfrak{g}$-module X is* **discretely decomposable** *if there is an increasing sequence of $\mathfrak{g}$-submodules of finite length:*

$$X = \bigcup_{m=0}^{\infty} X_m, \quad X_0 \subset X_1 \subset X_2 \subset \cdots . \tag{3.1}$$

We note that $\dim X_m = \infty$ *in most cases below.*

Next, consider the restriction of group representations.

Definition 3.2. *Let $G \supset G'$ be a pair of reductive Lie groups, and $\pi \in \widehat{G}$. We say that the restriction $\pi|_{G'}$ is* **infinitesimally discretely decomposable** *if the underlying $(\mathfrak{g}, K)$-module π_K is discretely decomposable as a $\mathfrak{g}'$-module.*

The terminology "discretely decomposable" is named after the following fact:

Theorem E ([Ko$_9$]). *Let (G, G') be a pair of reductive Lie groups, and π_K the underlying $(\mathfrak{g}, K)$-module of $\pi \in \widehat{G}$. Then* (i) *and* (ii) *are equivalent:*

i) *The restriction $\pi|_{G'}$ is infinitesimally discretely decomposable.*

ii) *The $(\mathfrak{g}, K)$-module π_K has a* **discrete branching law** *in the sense that π_K is isomorphic to an algebraic direct sum of irreducible $(\mathfrak{g}', K')$-modules.*

Moreover, the following theorem holds:

Theorem F (infinitesimal $\Rightarrow$ Hilbert space decomposition; [Ko$_{11}$]). *Let $\pi \in \widehat{G}$. If the restriction $\pi|_{G'}$ is infinitesimally discretely decomposable, then the restriction $\pi|_{G'}$ decomposes without continuous spectrum:*

$$\pi|_{G'} \simeq \sum_{\tau \in \widehat{G'}}^{\oplus} m_\pi(\tau)\tau \quad \textit{(a discrete direct sum of Hilbert spaces).} \tag{3.2}$$

At this stage, the multiplicity $m_\pi(\tau) := \dim \mathrm{Hom}_{G'}(\tau, \pi|_{G'})$ can be infinite.

However, for a reductive symmetric pair (G, G'), it is likely that the multiplicity of discrete spectrum is finite under the following assumptions, respectively.

(3.3) π is a discrete series representation for G.

(3.4) The restriction $\pi|_{G'}$ is infinitesimally discretely decomposable.

Conjecture 3.3 (Wallach, [X]). *$m_\pi(\tau) < \infty$ for any $\tau \in \widehat{G'}$ if (3.3) holds.*

Conjecture 3.4 ([Ko$_{11}$, Conjecture C]). *$m_\pi(\tau) < \infty$ for any $\tau \in \widehat{G'}$ if (3.4) holds.*

We note that Conjecture 3.4 for compact G' corresponds to Harish-Chandra's admissibility theorem. A first affirmative result for general non-compact G' was given in [Ko$_9$], which asserts that Conjecture 3.4 holds if π is attached to an elliptic coadjoint orbit. A special case of this assertion is:

Theorem G ([Ko$_9$]). *$m_\pi(\tau) < \infty$ for any $\tau \in \widehat{G'}$ if both (3.3) and (3.4) hold.*

In particular, Wallach's Conjecture 3.3 holds in the discretely decomposable case. We note that an analogous finite-multiplicity statement fails if **continuous spectrum** occurs in the restriction $\pi|_{G'}$ for a reductive symmetric pair (G, G'):

Counter Example 3.5 ([Ko$_{11}$]). $m_\pi(\tau)$ can be ∞ if neither (3.3) nor (3.4) holds.

Recently, I was informed by Huang and Vogan that they proved Conjecture 3.4 for any π [Hu-Vo].

A key step of Theorem G is to deduce the K'-admissibility of the restriction $\pi|_{K'}$ from the discreteness assumption (3.4), for which we employ Theorem H below. Let us explain it briefly. We write $\mathcal{V}_{\mathfrak{g}}(\pi)$ for the **associated variety** of the underlying $(\mathfrak{g}, K)$-module of π (see [Vo]), which is an algebraic variety contained in the **nilpotent cone** of $\mathfrak{g}^*$. Let $\mathrm{pr}_{\mathfrak{g}\to\mathfrak{g}'} : \mathfrak{g}^* \to (\mathfrak{g}')^*$ be the projection corresponding to $\mathfrak{g}' \subset \mathfrak{g}$. Here is a necessary condition for infinitesimal discrete decomposability:

Theorem H (criterion for discrete decomposability [Ko$_9$, Corollary 3.4]**).** *Let $\pi \in \widehat{G}$. If the restriction $\pi|_{G'}$ is infinitesimally discretely decomposable, then $\mathrm{pr}_{\mathfrak{g}\to\mathfrak{g}'}(\mathcal{V}_{\mathfrak{g}}(\pi))$ is contained in the nilpotent cone of $(\mathfrak{g}')^*$.*

We end this section with a useful information on irreducible summands.

Theorem I (size of irreducible summands, [Ko$_9$]**).** *Let $\pi \in \widehat{G}$. If the restriction $\pi|_{G'}$ is infinitesimally discretely decomposable, then any irreducible summand has the same associated variety, especially, the same Gelfand-Kirillov dimension.*

Here is a special case of Theorem I:

Example 3.6 (highest weight modules, [N-Oc-T]). Let G be the metaplectic group, and $G' = G'_1 G'_2$ is a dual pair with G'_1 compact. Let $\theta(\sigma)$ be an irreducible unitary highest weight module of G'_2 obtained as the **theta-correspondence** of $\sigma \in \widehat{G'_1}$. Then the associated variety of $\theta(\sigma)$ does not depend on σ, but only on G'_1.

An analogous statement to Theorem I fails if there exists continuous spectrum in the branching law $\pi|_{G'}$ (see [Ko$_{11}$] for counter examples).

4. Applications to representation theory

So far, we have explained basic theory of discretely decomposable restrictions of unitary representations for reductive Lie groups $G \supset G'$. Now, we ask what

discrete decomposability can do for representation theory. Let us clarify advantages of admissible restrictions, from which the following applications (and some more) have been brought out and seem to be promising furthermore.

1) Study of $\widehat{G'}$ as irreducible summands of $\pi|_{G'}$.
2) Study of $\widehat{G}$ by means of the restrictions to subgroups G'.
3) Branching laws of their own right.

4.1. From the view point of the study of $\widehat{G'}$ (smaller group), one of advantages of admissible restrictions is that each irreducible summand of the branching law $\pi|_{G'}$ gives an explicit construction of an element of $\widehat{G'}$.

Historically, an early success of this idea (in '70s and '80s) was the construction of irreducible highest weight modules (Howe, Kashiwara-Vergne, Adams, $\cdots$). A large part of these modules can be constructed as irreducible summands of discrete branching laws of the Weil representation (see Examples 2.3 (2) and 3.6).

This idea works also for non-highest weight modules. As one can observe from the criterion in Theorem B, the restriction $\pi|_{G'}$ tends to be discretely decomposable, if $\mathrm{AS}_K(\pi)$ is "small". In particular, if π is a **minimal representation** in the sense that its annihilator is the Joseph ideal, then a result of Vogan implies that $\mathrm{AS}_K(\pi)$ is one dimensional. Thus, there is a good possibility of finding subgroups G' such that $\pi|_{G'}$ is G'-admissible. This idea was used to construct "small" representations of subgroups G' by Gross-Wallach [Gr-W$_1$]. In the same line, discretely decomposable branching laws for non-compact G' are used also in the theory of automorphic forms for exceptional groups by J-S. Li [Li$_2$].

4.2. From the view point of the study of $\widehat{G}$ (larger group), one of advantages of admissible restrictions is to give a clue to a detailed study of representations of G by means of discrete branching laws.

Needless to say, an early success in this direction is the theory of $(\mathfrak{g}, K)$-modules (Lepowsky, Harish-Chandra, $\cdots$). The theory relies heavily on Harish-Chandra's admissibility theorem (Example 2.3 (1)) on the restriction of π to K.

Instead of a maximal compact subgroup K, this idea applied to a non-compact subgroup G' still works, especially in the study of "small" representations of G. In particular, this approach makes sense if the K-type structure is complicated but the G'-type structure is less complicated. Successful examples in this direction include:

1) To determine an explicit condition on λ such that a Zuckerman-Vogan module $A_{\mathfrak{q}}(\lambda)$ is non-zero, where we concern with the parameter λ outside the good range. In the setting of Example 2.6 (3), the author found in [Ko$_2$] a combinatorial formula on K_1-types of $A_{\mathfrak{q}}(\lambda)$ and determined explicitly when $A_{\mathfrak{q}}(\lambda) \neq 0$. The point here is that the computation of K-types of $A_{\mathfrak{q}}(\lambda)$ is too complicated to carry out because a lot of cancellation occurs in the generalized Blattner formula, while K_1-type formula (or G'-type formula for some non-compact subgroup G') behaves much simpler in this case.
2) To study a fine structure of standard representations. For example, Lee and Loke [Le-Lo] determined the Jordan-Hölder series and the unitarizability

of subquotients of certain degenerate non-unitary principal series representations π, by using G'-admissible restrictions for some non-compact reductive subgroup G'. Their method works successfully even in the case where K-type multiplicity of π is not one.

4.3. From the view point of finding explicit branching law, an advantage of admissible restrictions is that one can employ algebraic techniques because of the lack of continuous spectrum. A number of explicit branching laws are newly found (e.g. [D-Vs, Gr-$W_{1,2}$, Hu-P-S, $Ko_{1,3,4,8}$, Ko-$\emptyset_{1,2}$, Li_2, $Lo_{1,2}$, X]) in the context of admissible restrictions to non-compact reductive subgroups. A mysterious feature is that "different series" of irreducible representations may appear in discretely decomopsable branching laws (see [Ko_5, p.184] for a precise meaning), although all of them have the same Gelfand-Kirillov dimensions (Theorem I).

5. New discrete series for homogeneous spaces

Let $G \supset H$ be a pair of reductive Lie groups. Then, there is a G-invariant Borel measure on the homogeneous space G/H, and one can define naturally a unitary representation of G on the Hilbert space $L^2(G/H)$.

Definition 5.1. *We say π is a* **discrete series representation** *for G/H, if $\pi \in \widehat{G}$ is realized as a subrepresentation of $L^2(G/H)$.*

A discrete series representation corresponds to a discrete spectrum in the Plancherel formula for the homogeneous space G/H. One of basic problems in non-commutative harmonic analysis is:

Problem 5.2. 1) Find a condition on the pair of groups (G, H) such that there exists a discrete series representation for the homogeneous space G/H.

2) If exist, construct discrete series representations.

Even the first question has not found a final answer in the generality that (G, H) is a pair of reductive Lie groups. Here are some known cases:

Example 5.3. Flensted-Jensen, Matsuki and Oshima proved in '80s that discrete series representations for a reductive symmetric space G/H exist if and only if

$$\operatorname{rank} G/H = \operatorname{rank} K/(H \cap K). \tag{5.1}$$

This is a generalization of Harish-Chandra's condition, $\operatorname{rank} G = \operatorname{rank} K$, for a group manifold $G \times G/\operatorname{diag}(G) \simeq G$ ([FJ, Mk-Os]).

Our strategy to attack Problem 5.2 for more general (non-symmetric) homogeneous spaces G/H consists of two steps:

1) To embed G/H into a larger homogeneous space $\widetilde{G}/\widetilde{H}$, on which harmonic analysis is well-understood (e.g. symmetric spaces).
2) To take functions belonging to a discrete series representation $\mathcal{H}$ $(\hookrightarrow L^2(\widetilde{G}/\widetilde{H}))$, and to restrict them with respect to a submanifold G/H $(\hookrightarrow \widetilde{G}/\widetilde{H})$.

If G/H is "generic", namely, a **principal orbit** in $\widetilde{G}/\widetilde{H}$ in the sense of Richardson, then it is readily seen that discrete spectrum of the branching law $\pi|_G$ gives a discrete series for G/H ([Ko$_{10}$, §8]; see also [Hu, Ko$_{1,5}$, Li$_1$] for concrete examples).

However, some other interesting homogeneous spaces G/H occur as non-principal orbits on $\widetilde{G}/\widetilde{H}$, where the above strategy does not work in general. A remedy for this is to impose the **admissibility of the restriction** of π, which justifies the restriction of L^p-functions to submanifolds, and then gives rise to many non-symmetric homogeneous spaces that admit discrete series representations. For example, let us consider the case where $G = \widetilde{G}^\tau$ and $H = \widetilde{G}^\sigma$ for commuting involutive automorphisms τ and σ of $\widetilde{G}$ such that $\widetilde{G}/\widetilde{H}$ satisfies (5.1). Then by using Theorem C and an asymptotic estimate of invariant measures [Ko$_6$], we have:

Theorem J (discrete series for non-symmetric spaces, [Ko$_{10}$]). *Assume that there is $w \in W_\sigma$ such that*

$$\mathbb{R}_+ \text{-span}\, \Delta^+(\mathfrak{p})_{\sigma,w} \cap \sqrt{-1}(\mathfrak{t}_0^*)^{-\tau} = \{0\}. \tag{5.2}$$

Then there exist infinitely many discrete series representations for any homogeneous space of G that goes through $x\widetilde{H} \in \widetilde{G}/\widetilde{H}$ for any $x \in \widetilde{K}$.

We refer to [Ko$_{10}$, Theorem 5.1] for definitions of a finite group W_σ and $\Delta^+(\mathfrak{p})_{\sigma,w}$. The point here is that the condition (5.2) can be easily checked.

For instance, if $G \simeq Sp(2n,\mathbb{R}) \simeq \widetilde{G}/\widetilde{H}$ (a group manifold), then Theorem J implies that there exist discrete series on all homogeneous spaces of the form:

$$G/H = Sp(2n,\mathbb{R})/(Sp(n_0,\mathbb{C}) \times GL(n_1,\mathbb{C}) \times \cdots \times GL(n_k,\mathbb{C})), \quad (\sum n_i = n).$$

The choice of x in Theorem J corresponds to the partition $(n_0, n_1, \ldots, n_k)$. We note that the above G/H is a symmetric space if and only if $n_1 = n_2 = \cdots = n_k = 0$.

The restriction of unitary representations gives new methods even for symmetric spaces where harmonic analysis has a long history of research. Let us state two results that are proved by the theory of discretely decomposable restrictions.

Theorem K (holomorphic discrete series for symmetric spaces). *Suppose G/H is a non-compact irreducible symmetric space. Then* (i) *and* (ii) *are equivalent:*

i) *There exist unitary highest weight representations of G that can be realized as subrepresentations of $L^2(G/H)$.*
ii) *G/K is Hermitian symmetric and $H/(H \cap K)$ is its totally real submanifold.*

This theorem in the group manifold case is a restatement of Harish-Chandra's well-known result. The implication (ii) $\Rightarrow$ (i) was previously obtained by a different geometric approach ('Olafsson-Ørsted [Ol-Ø]). Our proof uses a general theory of discretely decomposable restrictions, especially, Theorems B, H and J.

Theorem L (exclusive law of discrete spectrum for restriction and induction). *Let G/G' be a non-compact irreducible symmetric space, and $\pi \in \widehat{G}$. Then both (1) and (2) cannot occur simultaneously.*

1) *The restriction $\pi|_{G'}$ is infinitesimally discretely decomposable.*
2) *π is a discrete series representation for the homogeneous space G/G'.*

We illustrate Theorems K and L by $G = SL(2,\mathbb{R})$. The examples below are well-known results on harmonic analysis, however, the point is that they can be proved by a simple idea coming from restrictions of unitary representations.

Example 5.4. 1) Holomorphic discrete series exist for $G/H = SL(2,\mathbb{R})/SO(1,1)$ (a hyperboloid of one sheet). This is explained by Theorem K because the geodesic $H/(H \cap K)$ is obviously totally real in the Poincaré disk $G/K = SL(2,\mathbb{R})/SO(2)$.
2) There is no discrete series for the Poincaré disk $G/K = SL(2,\mathbb{R})/SO(2)$. This fact is explained by Theorem L because any representation of G is obviously discretely decomposable when restricted to a compact K.

6. Modular varieties, vanishing theorem

Retain the setting as in Definition 2.2. Let $\Gamma' \subset \Gamma$ be cocompact torsion-free discrete subgroups of $G' \subset G$, respectively. For simplicity, let G' be a semisimple Lie group without compact factors. Then, both of the double cosets $X := \Gamma\backslash G/K$ and $Y := \Gamma'\backslash G'/K'$ are compact, orientable, locally Riemannian symmetric spaces. Then, the inclusion $G' \hookrightarrow G$ induces a natural map $\iota : Y \to X$. The image $\iota(Y)$ defines a totally geodesic submanifold in X. Consider the induced homomorphism of the homology groups of degree $m := \dim Y$,

$$\iota_* : H_m(Y;\mathbb{Z}) \to H_m(X;\mathbb{Z}).$$

The **modular symbol** is defined to be the image $\iota_*[Y] \in H_m(X;\mathbb{Z})$ of the fundamental class $[Y] \in H_m(Y;\mathbb{Z})$. Though its definition is simple, the understanding of modular symbols is highly non-trivial.

Let us first recall some results of Matsushima-Murakami and Borel-Wallach on the de Rham cohomology group $H^*(X;\mathbb{C})$ summarized as:

$$H^*(X;\mathbb{C}) = \bigoplus_{\pi \in \widehat{G}} H^*(X)_\pi, \quad H^*(X)_\pi := \mathrm{Hom}_G(\pi, L^2(\Gamma\backslash G)) \otimes H^*(\mathfrak{g}, K; \pi_K). \quad (6.1)$$

The above result describes the topology of a single X by means of representation theory. For the topology of the pair (Y, X), we need restrictions of representations:

Theorem M (vanishing theorem for modular symbols, [Ko-Od]**).** *If*

$$\mathrm{AS}_K(\pi) \cap \mathrm{Cone}(G') = \{0\}, \quad \pi \neq \mathbf{1},$$

then the modular symbol $\iota_*[Y]$ *is annihilated by the* π*-component* $H^m(X)_\pi$ *in the perfect paring* $H^m(X;\mathbb{C}) \times H_m(X;\mathbb{C}) \to \mathbb{C}$.

Theorem M determines, for example, the middle Hodge components of totally real modular symbols of compact Clifford-Klein forms of type IV domains.

The discreteness of irreducible decompositions plays a crucial role both in Matsushima-Murakami's formula (6.1) and in a vanishing theorem for modular varieties (Theorem M). In the former $L^2(\Gamma\backslash G)$ is G-admissible (Gelfand and Piateski-Shapiro), while the restriction $\pi|_{G'}$ is G'-admissible (cf. Theorem B) in the latter.

References

[A] M. F. Atiyah, *The Harish-Chandra character*, London Math. Soc. Lecture Note Series **34** (1979), 176–181.

[D-Vs] M. Duflo and J. Vargas, in preparation.

[FJ] M. Flensted-Jensen, *Discrete series for semisimple symmetric spaces*, Annals of Math. **111** (1980), 253–311.

[Ge-Gv] I. M. Gelfand and M. I. Graev, *Geometry of homogeneous spaces, representations of groups in homogeneous spaces, and related questions of integral geometry*, Transl. II. Ser., A. M. S. **37** (1964), 351–429.

[Gr-W$_1$] B. Gross and N. Wallach, *A distinguished family of unitary representations for the exceptional groups of real rank* $= 4$, Progress in Math. **123** (1994), Birkhäuser, 289–304.

[Gr-W$_2$] B. Gross and N. Wallach, *Restriction of small discrete series representations to symmetric subgroups*, Proc. Sympos. Pure Math. **68** (2000), A.M.S., 255–272.

[Ho$_1$] R. Howe, *θ-series and invariant theory*, Proc. Sympos. Pure Math. **33** (1979), A.M.S., 275–285.

[Ho$_2$] R. Howe, *Wave front sets of representations of Lie groups*, Automorphic forms, representation theory, and arithmetic (1981), Tata, 117–140.

[Hu] J-S. Huang, *Harmonic analysis on compact polar homogeneous spaces*, Pacific J. Math. **175** (1996), 553—569.

[Hu-P-S] J-S. Huang, P. Pandžić, and G. Savin, *New dual pair correspondences*, Duke Math. **82** (1996), 447–471.

[Hu-Vo] J-S. Huang and D. Vogan, personal communications (2001).

[J-Vr] H. P. Jakobsen and M. Vergne, *Restrictions and expansions of holomorphic representations*, J. Funct. Anal. **34** (1979), 29–53.

[Ks-Vr] M. Kashiwara and M. Vergne, *K-types and singular spectrum*, Lect. Notes in Math., vol. 728, Springer, 1979, 177–200.

[Kn-Vo] A. Knapp and D. Vogan, Jr., *Cohomological Induction and Unitary Representations*, Princeton U.P., 1995.

[Ko$_1$] T. Kobayashi, *Unitary representations realized in L^2-sections of vector bundles over semi-simple symmetric spaces*, Proc. of the 27-28th Symp. of Funct. Anal. and Real Anal. (1989), Math. Soc. Japan, 39–54. (Japanese)

[Ko_2] T. Kobayashi, *Singular Unitary Representations and Discrete Series for Indefinite Stiefel Manifolds* $U(p,q;\mathbb{F})/U(p-m,q;\mathbb{F})$, Memoirs of A.M.S., vol. 462, 1992.

[Ko_3] T. Kobayashi, *The Restriction of* $A_{\mathfrak{q}}(\lambda)$ *to reductive subgroups*, Part I, Proc. Japan Acad. **69** (1993), 262–267; Part II, ibid. **71** 1995, 24–26.

[Ko_4] T. Kobayashi, *Harmonic analysis on homogeneous manifolds of reductive type and unitary representation theory*, Transl., Series II, Selected Papers on Harmonic Analysis, Groups, and Invariants **183** (1998), A.M.S., 1–31.

[Ko_5] T. Kobayashi, *Discrete decomposability of the restriction of* $A_{\mathfrak{q}}(\lambda)$ *with respect to reductive subgroups and its applications*, Invent. Math. **117** (1994), 181–205.

[Ko_6] T. Kobayashi, *Invariant measures on homogeneous manifolds of reductive type*, J. reine und angew. Math. **490** (1997), 37–53.

[Ko_7] T. Kobayashi, *Discrete decomposability of the restriction of* $A_{\mathfrak{q}}(\lambda)$ *with respect to reductive subgroups* II — *micro-local analysis and asymptotic* K*-support*, Annals of Math. **147** (1998), 709–729.

[Ko_8] T. Kobayashi, *Multiplicity free branching laws for unitary highest weight modules*, Proceedings of the Symposium on Representation Theory held at Saga, Kyushu (K. Mimachi, ed.), 1997, 7–13.

[Ko_9] T. Kobayashi, *Discrete decomposability of the restriction of* $A_{\mathfrak{q}}(\lambda)$ *with respect to reductive subgroups* III — *restriction of Harish-Chandra modules and associated varieties*, Invent. Math. **131** (1998), 229–256.

[Ko_{10}] T. Kobayashi, *Discrete series representations for the orbit spaces arising from two involutions of real reductive Lie groups*, J. Funct. Anal. **152** (1998), 100–135.

[Ko_{11}] T. Kobayashi, *Discretely decomposable restrictions of unitary representations of reductive Lie groups —examples and conjectures*, Adv. Stud. Pure Math. **26** (2000), 99-127.

[Ko_{12}] T. Kobayashi, *Theory of discrete decomposable branching laws of unitary representations of semisimple Lie groups and some applications*, Sugaku Exposition, Transl. Ser., A.M.S. (to appear).

[Ko_{13}] T. Kobayashi, in preparation.

[Ko-Od] T. Kobayashi and T. Oda, *Vanishing theorem of modular symbols on locally symmetric spaces*, Comment. Math. Helvetici **73** (1998), 45–70.

[Ko-$\mathrm{Ø}_1$] T. Kobayashi and B. Ørsted, *Conformal geometry and branching laws for unitary representations attached to minimal nilpotent orbits*, C. R. Acad. Sci. Paris **326** (1998), 925–930.

[Ko-$\mathrm{Ø}_2$] T. Kobayashi and B. Ørsted, *Analysis on the minimal representation of* $O(p,q)$, I, II, III, preprint.

[Le-Lo] S-T. Lee and H-Y. Loke, *Degenerate principal series of* $U(p,q)$ *and* $Spin(p,q)$, preprint.

[Li_1] J-S. Li, *On the discrete series of generalized Stiefel manifolds*, Trans. A.M.S. **340** (1993), 753–766.

[Li_2] J-S. Li, *Two reductive dual pairs in groups of type E*, Manuscripta Math. **91** (1996), 163–177.

[Lo_1] H-Y. Loke, *Restrictions of quaternionic representations*, J. Funct. Anal. **172** (2000), 377–403.

[Lo_2] H-Y. Loke, *Howe quotients of unitary characters and unitary lowest weight modules*, preprint.

[Mg] G. Margulis, *Existence of compact quotients of homogeneous spaces, measurably proper actions, and decay of matrix coefficients*, Bul. Soc. Math. France **125** (1997), 1–10.

[Mt] S. Martens, *The characters of the holomorphic discrete series*, Proc. Nat. Acad. Sci. USA **72** (1975), 3275-3276.

[Mk-Os] T. Matsuki and T. Oshima, *A description of discrete series for semisimple symmetric spaces*, Adv. Stud. Pure Math. **4** (1984), 331–390.

[N-Oc-T] K. Nishiyama, H. Ochiai, and K. Taniguchi, *Bernstein degree and associated cycles of Harish-Chandra modules — Hermitian symmetric case*, Asterisque **273** (2001), 13–80.

[Oh] H. Oh, *Tempered subgroups and representations with minimal decay of matrix coefficients*, Bull. Soc. Math. France **126** (1998), 355–380.

[Ol-Ø] G. Ólafsson and B. Ørsted, *The holomorphic discrete series of an affine symmetric space,* I, J. Funct. Anal. **81** (1988), 126–159.

[Ø-Vs] B. Ørsted and J. Vargas, *Restriction of square integrable representations: discrete spectrum*, preprint.

[Vo] D. Vogan, Jr., *Associated varieties and unipotent representations*, Progress in Math. **101** (1991), Birkhäuser, 315–388.

[Vo-Z] D. Vogan, Jr. and G. Zuckerman, *Unitary representations with non-zero cohomology*, Compositio Math. **53** (1984), 51–90.

[X] J. Xie, *Restriction of discrete series of* $SU(2,1)$ *to* $S(U(1) \times U(1,1))$, J. Funct. Anal. **122** (1994), 478–518, ph.D. dissertation, Rutgers University.

ICM 2002 · Vol. II · 629–635

Representations of Algebraic Groups and Principal Bundles on Algebraic Varieties

Vikram Bhagvandas Mehta*

Abstract

In this talk we discuss the relations between representations of algebraic groups and principal bundles on algebraic varieties, especially in characteristic p. We quickly review the notions of stable and semistable vector bundles and principal G-bundles , where G is any semisimple group. We define the notion of a low height representation in characteristic p and outline a proof of the theorem that a bundle induced from a semistable bundle by a low height representation is again semistable. We include applications of this result to the following questions in characteristic p:

1) Existence of the moduli spaces of semistable G-bundles on curves.

2) Rationality of the canonical parabolic for nonsemistable principal bundles on curves.

3) Luna's etale slice theorem.

We outline an application of a recent result of Hashimoto to study the singularities of the moduli spaces in (1) above, as well as when these spaces specialize correctly from characteristic 0 to characteristic p. We also discuss the results of Laszlo-Beauville-Sorger and Kumar-Narasimhan on the Picard group of these spaces. This is combined with the work of Hara and Srinivas-Mehta to show that these moduli spaces are F-split for p very large. We conclude by listing some open problems, in particular the problem of refining the bounds on the primes involved.

2000 Mathematics Subject Classification: 22E46, 14D20.
Keywords and Phrases: Semistable bundles, Low-height representations.

1. Some Definitions

We begin with some basic definitions:

Let V be a vector bundle on a smooth projective curve X of genus g over an algebraically closed field (in any characteristic).

*Tata Institute of Fundamental Research, Mumbai, India. E-mail: vikram@math.tifr.res.in

Definition 1.1: *V is* stable *(respectively* semi-stable*) if for all subbundles W of V, we have*

$$\mu(W) \stackrel{=}{def} \deg W/rk\ W \ < (\leq)$$

$$\mu(V) \stackrel{=}{def} \deg V/rk\ V.$$

For integers r and d with $r > 0$, one constructs the moduli spaces $U^s(r,d)(U(r,d))$ of stable (semistable) vector bundles of rank r and degree d, using Geometric Invariant Theory (G.I.T.).

If the ground field is $\mathbb{C}$, the complex numbers, one has the basic (genus $X \geq 2$):

Theorem 1.2: *Let V have degree 0. Then V is stable $\Leftrightarrow$ $V \simeq V_\sigma$, for some irreducible representation $\sigma : \pi_1(X) \to U(n)$.*

This is due to Narasimhan-Seshadri. Note that $H \to X$ is a principal $\pi_1(X)$ fibration, where H is the upper-half plane. Any $\sigma : \pi_1(X) \to GL(n,\mathbb{C})$ gives a vector bundle of rank n on $X, V_\sigma = H \times^{\pi_1(X)} \mathbb{C}^n$.

Remark 1.3: It follows from Theorem 1.2 that if V is a semistable bundle on a curve X over $\mathbb{C}$, then $\otimes^n(V), S^n(V)$, in fact any bundle induced from V is again semistable. By Lefschetz, this holds for any algebraically closed field of characteristic 0.

Remark 1.4: In general, a subbundle W of a vector bundle V is a reduction of the structure group of the principal bundle of V to a maximal parabolic of $GL(n), n = rank\ V$. This is in turn equivalent to a section σ of the associated fibre space:

$$E \times^{GL(n)} \ GL(n)/P.$$

Now let X be a smooth curve and $E \stackrel{\pi}{\to} X$ a principal G-bundle on X, where G is a semisimple (or even a reductive) group in any characteristic.

Definition 1.5: E is stable (semistable) $\Leftrightarrow \forall$ maximal parabolics P of $G, \forall$ sections σ of $E(G/P)$, we have degree $\sigma^\# T_\pi > 0 (\geq 0)$, where T_π is the relative tangent bundle of $E(G/P) \stackrel{\pi}{\to} X$.

Over $\mathbb{C}$, we have the following [18]:

Theorem 1.6: *$E \to X$ is stable $\Leftrightarrow$ $E \simeq E_\sigma$ for some irreducible representation $\sigma : \pi_1(X) \to K$, the maximal compact of G.*

The analogue of Remark 1.3 is valid in this general situation.

Remark 1.7: One can analogously define stable and semistable vector bundles and principal bundles on normal projective varieties of dimension > 1. Again, in characteristic 0, bundles induced from semistable bundles continue to be semistable.

Remark 1.8: In characteristic p, bundles induced from semistable bundles need not be semistable, in general[7]. In this lecture we shall examine some conditions when this does hold, and also discuss some applications to the moduli spaces of principal G-bundles on curves.

2. Low height representations

Here we introduce the basic notion of a low height representation in characteristic p. Let $f : G \to SL(n) = SL(V)$ be a representation of G in char p, G being reductive. Fix a Borel B and a Torus T in G. Let $L(\lambda_i), 1 \leq i \leq m$, be the simple G-modules occurring in the Jordan-Holder filtration of V. Write each λ_i as $\sum_j q_{ij}\alpha_j$, where $\{\alpha_j\}$ is the system of simple roots corresponding to B and $q_{ij} \in Q\ \forall i, j$. Define $ht\lambda_i = \sum_j q_{ij}$. Then one has the basic [9,20]:

Definition 2.1: *f is a low-height representation of G, or V is a low-height module over G, if $2ht(\lambda_i) < p\ \forall i$.*

Remark 2.2: If $2ht(\lambda_i) < p\ \forall i$, then it easily follows that V is a completely reducible G-module. In fact for any subgroup Γ of G, V is completely reducible over $\Gamma \Leftrightarrow \Gamma$ itself is completely reducible in G. By definition, an abstract subgroup Γ of G is completely reducible in $G \Leftrightarrow$ for any parabolic P of G, if Γ is contained in P then Γ is contained in a Levi component L of P. These results were proved by Serre[20] using the notion of a saturated subgroup of G.

In general, denote sup $(2ht\ \lambda_i)$ by $ht_G V$. If V is the standard $SL(n)$ module, then $ht_{SL(n)} \wedge^i (V) = i(n-i), 1 \leq i \leq n-1$. More generally, $ht_G(V_1 \otimes V_2) = ht_G V_1 + ht_G V_2$. The following theorem is the key link between low-height representations and semistability of induced bundles [9]:

Theorem 2.3: *Let $E \to X$ be a semistable G-bundle, where G is semisimple and the base X is a normal projective variety. Let $f : G \to SL(n)$ be a low-height representation. Then the induced bundle $E(SL(n))$ is again semistable.*

The proof is an interplay between the results of Bogomolov, Kempf, Rousseau and Kirwan in G.I.T. on one hand and the results of Serre mentioned earlier on the other. The group scheme $E(G)$ over X acts on $E(SL(n)/P)$ and assume that σ is a section of the latter. Consider the generic point K of X and its algebraic closure $\overline{K}$. Then $E(G)_{\overline{K}}$ acts on $E(SL(n)/P)_{\overline{K}}$, and σ is a K-rational point of the latter. There are 2 possibilities:

1) σ is $G.I.T$ semistable. In this case, one can easily prove that deg $\sigma^{\#} T_\pi \geq 0$.
2) σ is $G.I.T.$ unstable, i.e., not semistable. Let $P(\sigma)$ be the Kempf-Rousseau parabolic for σ, which is defined over $\overline{K}$. For deg $\sigma^{\#} T_\pi$ to be ≥ 0 it is *sufficient* that $P(\sigma)$ is defined over K. Note that since V is a low-height representation of G, one has $p \geq h$. One then has ([20]).

Proposition 2.4: *If $p \geq h$, there is a unique G-invariant isomorphism log: $G^u \to \underline{g}_{\text{nilp}}$, where G^u is the unipotent variety of G and $\underline{g}_{\text{nilp}}$ is the nilpotent variety of $\underline{g} = Lie\ G$.*

Proposition 2.4 is used in

Proposition 2.5: *Let H be any semisimple group and W a low-height representation of H. Let $W_1 \subset W$ and assume that $\exists X \in Lie\ H$, X nilpotent such that $X \in Lie\ (Stab\ (W_1))$. Then in fact one has $X \in Lie\ [Stab\ (W_1)_{\text{red}}]$.*

Along with some facts from $G.I.T$, Proposition 2.5 enables us to prove that $P(\sigma)$ is in fact defined over K, thus finishing the sketch of the proof of Theorem 2.3. See also Ramanathan-Ramanan [19]. One application of low-height representations

is in the proof of a conjecture of Behrend on the rationality of the canonical parabolic or the instability parabolic. If V is a nonsemistable bundle on a variety X, then one can show that there exists a flag $V^{\cdot}$,

$$0 = V_0 \subset V_1 \subset V_2 \cdots \subset V_n = V$$

of subbundles of V with the properties:

(1) Each V_i/V_{i-1} is semistable and $\mu\ (V_i/V_{i-1}) > \mu\ (V_{i+1}/V_i), 1 \leq i \leq n-1$.
(2) The flag $V^{\cdot}$ as in (1) is unique and infinitesimally unique, i.e., $V^{\cdot}$ is defined over any field of definition of X and V. Such a flag corresponds to a reduction to a parabolic P of $GL(n)$ and properties (1) and (2) may be expressed as follows: the *elementary* vector bundles on X associated to P all have positive degree and $H^0(X, E(\underline{g})/E(\underline{p})) = 0$, where $\underline{g} = \text{Lie } GL(n)$ and $\underline{p} = \text{Lie } P$.

One may ask whether there is a such a canonical reduction for a nonsemistable principal G bundle $E \to X$. Such a reduction was first asserted first by Ramanathan [18], and then by Atiyah-Bott[1] ,both over $\mathbb{C}$ and both without proofs. It was Behrend [5], who first proved the existence and uniqueness of the canonical reduction to the instability parabolic in all characteristics. Further, Behrend conjectured that $H^0(X, E(\underline{g})/E(\underline{p})) = 0$.

In characteristic zero, one can check that all three definitions of the instability parabolic coincide and that Behrend's conjecture is valid. In characteristic p, one uses low-height representations to show the equality of the three definitions and prove Behrend's conjecture [14].

Theorem 2.6: *Let $E \to X$ be a nonsemistable principal G-bundle in char p. Assume that $p > 2dimG$. Then all the 3 definitions coincide and further we have $H^0(X, E(\underline{g})/E(\underline{p})) = 0$, where $\underline{p} = Lie\ P$ and P is the instability parabolic.*

Theorem 2.6 is useful, among other things, for classifying principal G-bundles on $\mathbb{P}^1$ and $\mathbb{P}^2$ in characteristic p.

If V is a finite-dimensional representation of a semisimple group G (in any characteristic), then the G.I.T. quotient $V//G$ parametrizes the closed orbits in V. Now, let the characteristic be zero and let $v_0 \in V$ have a closed orbit. Then Luna's étale slice theorem says that $\exists$ a locally closed non-singular subvariety S of V such that $v_0 \in S$ and $S//G_{v_o}$ is isomorphic to $V//G$, locally at v_0, in the étale topology. Here G_{v_0} is the stabilizer of v_0. The proof uses the fact that G_{v_0} is a reductive subgroup of G (not necessarily connected!), hence V is a completely reducible G module. In characteristic p, one has to assume that V is a low-ht representation of G. Then the conclusion of Luna's étale slice theorem is still valid: to be more precise, let V be a low-ht representation of G and let $v_0 \in V$ have a closed orbit. Put $H = \text{Stab } (v_0)$. The essential point, as in characteristic 0, is to prove the complete reducibility of V over H. Using the low-ht assumption, one shows that every $X \in \text{Lie } H$ with X nilpotent can be integrated to a homomorphism $G_a \to H$ with tangent vector X. Now, under the hypothesis of low-ht, one shows that H_{red} is a saturated subgroup of G and $(H_{\text{red}} : H^0_{\text{red}})$ is prime to p. This shows that V

is a completely reducible H_{red} module. Further, one shows that H_{red} is a normal subgroup of H with H/H_{red} a finite group of multiplicative type, i.e. a finite subgroup of a torus. Now the complete reducibility of V over H follows easily [11]. Just as in characteristic zero, one deduces the existence of a smooth H-invariant subvariety S of V such that $v_0 \in S$ and $S//H$ is locally isomorphic to $V//G$ at v_0. This result is used in the construction of the moduli space M_G to be described in the next section.

3. Construction of the moduli spaces

The moduli spaces of semistable G-bundles on curves were first constructed by Ramanathan over $\mathbb{C}$ [16,17], then by Faltings and Balaji-Seshadri in characteristic 0 [3,6]. There are 3 main points in Ramanathan's construction:

1. If $E \to X$ is semistable, then the adjoint bundle $E(\underline{g})$ is semistable.
2. If $E \to X$ is polystable, then $E(\underline{g})$ is also polystable.
3. A semisimple Lie Algebra in char 0 is rigid.

The construction of M_G in char p was carried out in [2,15]. We describe the method of [15] first : points (1) and (2) are handled by Theorem 2.3 and the following [11] :

Theorem 3.1: *Let $E \to X$ be a polystable G-bundle over a curve in char p. Let $\sigma : G \to SL(n) = SL(V)$ be a representation such that all the exterior powers $\wedge^i V, 1 \leq i \leq n-1$, are low-height representations. Then the induced bundle $E(V)$ is also polystable.*

The proof uses Luna's étale slice theorem in char p and Theorem 2.3.

Now one takes a total family T of semistable G bundle on X and takes the good quotient of T to obtain M_G in char p. Theorem 3.1 is used to identify the closed points of M_G as the isomorphism classes of polystable G-bundles, just as in char 0. The semistable reduction theorem is proved by lifting to characteristic 0 and then applying Ramanathan's proof (in which (3) above plays a crucial role). This construction follows Ramanathan very closely and, as is clear, one has to make low-height assumptions as in Theorem 3.1.

The method of [2] follows the one in [3] with some technical and conceptual changes. One chooses an embedding $G \to SL(n)$ and a representation W for $SL(n)$ such that (1) G is the stabilizer of some $w_0 \in W$. (2) W is a "low separable index representation" of $SL(n)$, i.e., all stabilizers are reduced and W is low-height over $SL(n)$. The semistable reduction theorem is proved using the theory of Bruhat-Tits. Here also suitable low-height assumptions have to be made.

4. Singularities and specialization of the moduli spaces

We first discuss the singularities of M_G, assuming throughout that G is simply connected. In char 0, M_G has rational singularities, this follows from Boutot's theorem. In char p, the following theorem due to Hashimoto [8] is relevant:

Theorem 4.1: *Let V be a representation of G such that all the symmetric powers $S^n(V)$ have a good filtration. Then the ring of invariant $[S^{\cdot}(V)]^G$ is strongly F-regular.*

Strongly F-regularity is a notion in the theory of tight closure in commutative algebra. We just note that if a geometric domain is strongly F-regular then it is normal,Cohen-Macaulay, F-split and has "rational-like" singularities. Now let $t \in M_G$ be the "worst point", i.e., the trivial G-bundle on X.

The local ring $(\mathcal{O}_{M_G}, t)^\wedge$ is isomorphic to $(S^{\cdot}(W)//G)^\wedge$, where W = direct sum of g copies of $\underline{g}$, with G acting diagonally. If p is a good prime for G , then Hashimoto's theorem implies that $\mathcal{O}_{M_G,t}$ is strongly F-regular. The other points of M_G are not so well understood. This would require a detailed study of the automorphism groups of polystable bundles, both in char 0 and p, and of their invariants of the slice representations. This is necessary also to study the specialization problem, i.e., when M_G in char 0 specializes to M_G in char p. One has to show that the invariants of the slice representations in char 0 specialize to the invariants in char p. However for G=SL(n),the situation is much simpler. One can write down the automorphism group of a polystable bundle and its representation on the local moduli space. Consequently, one expects the moduli spaces to specialize correctly and that the local rings of M_G are strongly F-regular in all positive characteristics.

We briefly discuss Pic M_G in char 0. It follows from [4,10] that M_G has the following properties in char 0:

1. Pic $M_G \simeq \mathbb{Z}$.
2. M_G is a normal projective, Gorenstein variety with rational singularities and with K negative ample.

Now let X be a normal,Cohen-Macaulay variety in char 0. It is proved in [13],in response to a conjecture of Karen Smith, that if X has rational singularities, then the reduction of X mod p is F-rational for all large p. This result together with 1 and 2 above imply that M_G reduced mod p is F-split for all large p. We cannot give effective bounds on the primes involved. One partial result is known in this direction[12].

Acknowledgement: I would like to thank my colleagues S. Ilangovan, A.J. Parameswaran and S. Subramanian for their help in preparing this report and T.T. Nayya and H for constant help and encouragement.

References

[1] M.F. Atiyah, R. Bott, The Yang-Mills equations over Riemann Surfaces, *Phil. Trans. R. Soc.* London A 308 (1982), 523–615.

[2] V. Balaji, A.J. Parameswaran Semistable Principal Bundles-II (in positive characteristics) to appear in *Transformation Groups.*

[3] V. Balaji, C.S. Seshadri Semistable Principal Bundles-I (in characteristic zero), to appear in *Journal of Algebra.*

[4] A. Beauville, Y. Laszlo, C. Sorger, The Picard Group of the moduli of G-bundles over curves, *Compositio Math.* 112, (1998), No.2, 183–216.

[5] K. Behrend, Semistability of reductive group schemes over curves, *Math. Ann.* 301 (1995), 281–305.
[6] G. Faltings, Stable G-bundles and projective connections, *J. Algebraic Geom.* 2 (1993) No.3, 507–568.
[7] D. Gieseker, Stable Vector Bundles and the Frobenius morphism, *Ann. Sci. Ecol. Nor. Sup.* 6, (1973).
[8] M. Hashimoto, Good filtrations of symmetric algebras and strong F-regularity of invariant subrings, *Math. Z.* 236 (2001), No.3, 605–623.
[9] S. Ilangovan, V.B. Mehta, A.J. Parameswaran, Semistability and Semisimplicity in representations of low-height in positive characteristics, preprint.
[10] S. Kumar, M.S. Narasimhan, Picard group of the moduli spaces of G-bundles, *Math. Ann.* 308, (1997), No.1, 155–173.
[11] V.B. Mehta, A.J. Parameswaran, Geometry of low-height representations, *Proceedings of the International Colloquium on Algebra, Arithmetic and Geometry*, (ed. R. Parimala), TIFR Mumbai 2000.
[12] V.B. Mehta, T.R. Ramadas, Moduli of vector bundles, Frobenius splitting and invariant theory, *Ann. of Math.* (2) 144, (1996), 269–313.
[13] V.B. Mehta, V. Srinivas, A characterization of rational singularities, *Asian J. Math.*, Vol.1, (1997), No.2, 249–271,.
[14] V.B. Mehta, S. Subramanian, On the Harder-Narasimhan Filtration of Principal Bundles, *Proceedings of the International Colloquium on Algebra, Arithmetic and Geometry*, (ed. R. Parimala), TIFR Mumbai 2000.
[15] V.B. Mehta, S. Subramanian, Moduli of Principal G-bundles on curves in positive characteristic, in preparation.
[16] A. Ramanathan, Moduli for principal bundles over algebraic curves I, *Proc. Indian Acad. Sci. Math. Sci.*, 106, (1996), No.3, 301–328.
[17] A. Ramanathan, Moduli for principal bundles over algebraic curves II, *Proc. Indian Acad. Sci. Math. Sci.*, 106 (1996), No.4, 421–449.
[18] A. Ramanathan, Moduli for principal bundles, in: *Algebraic Geometry*, Proceedings, Copenhagen 1978, 527–533, Lecture Notes in Mathematics vol. 732, Springer.
[19] S. Ramanan, A. Ramanathan, Some remarks on the instability flag, *Tohoku Math. Journal* 36, (1984), 269–291.
[20] J-P. Serre, *Moursund Lectures*, University of Oregon Mathematics Department, notes by W.E. Duckworth (1998).

ICM 2002 · Vol. II · 637–642

Clifford Algebras and the Duflo Isomorphism

E. Meinrenken*

Abstract

This article summarizes joint work with A. Alekseev (Geneva) on the Duflo isomorphism for quadratic Lie algebras. We describe a certain quantization map for Weil algebras, generalizing both the Duflo map and the quantization map for Clifford algebras. In this context, Duflo's theorem generalizes to a statement in equivariant cohomology.

2000 Mathematics Subject Classification: 17B, 22E60, 15A66, 55N91.
Keywords and Phrases: Clifford algebras, Quadratic Lie algebras, Duflo map, Equivariant cohomology.

1. Introduction

The universal enveloping algebra $U(\mathfrak{g})$ of a Lie algebra $(\mathfrak{g}, [\cdot,\cdot]_{\mathfrak{g}})$ is the quotient of the tensor algebra $T(\mathfrak{g})$ by the relations, $\xi\xi' - \xi'\xi = [\xi,\xi']_{\mathfrak{g}}$. The inclusion of the symmetric algebra $S(\mathfrak{g})$ into $T(\mathfrak{g})$ as totally symmetric tensors, followed by the quotient map, gives an isomorphism of $\mathfrak{g}$-modules

$$\mathrm{sym} : S(\mathfrak{g}) \to U(\mathfrak{g}) \tag{1.1}$$

called the *symmetrization map.* The restriction of sym to $\mathfrak{g}$-invariants is a vector space isomorphism, but not an algebra isomorphism, from invariant polynomials to the center of the enveloping algebra. Let $J \in C^\infty(\mathfrak{g})$ be the function

$$J(\xi) = \det(j(\mathrm{ad}_\xi)), \quad j(z) = \frac{\sinh(z/2)}{z/2},$$

and $J^{1/2}$ its square root (defined in a neighborhood of $\xi = 0$). Denote by $\widehat{J^{1/2}}$ the infinite order differential operator on $S\mathfrak{g} \subset C^\infty(\mathfrak{g}^*)$, obtained by replacing the

*Department of Mathematics, University of Toronto, 100 St. George Street, Toronto, ON M6R1G7, Canada. E-mail: mein@math.toronto.edu

variable $\xi \in \mathfrak{g}$ with a directional derivative $\frac{\partial}{\partial \mu}$, where μ is the dual variable on $\mathfrak{g}^*$. Duflo's celebrated theorem says that the composition

$$\mathrm{sym} \circ \widehat{J^{1/2}} : \ S\mathfrak{g} \to U(\mathfrak{g})$$

restricts to an algebra isomorphism, $(S\mathfrak{g})^{\mathfrak{g}} \to \mathrm{Cent}(U(\mathfrak{g}))$. In more geometric language, Duflo's theorem gives an isomorphism between the algebra of invariant constant coefficient differential operators on $\mathfrak{g}$ and bi-invariant differential operators on the corresponding Lie group G.

The purpose of this note is to give a quick overview of joint work with A. Alekseev [1, 2], in which we obtained a new proof and a generalization of Duflo's theorem for the special case of a *quadratic* Lie algebra. That is, we assume that $\mathfrak{g}$ comes equipped with an invariant, non-degenerate, symmetric bilinear form B. Examples of quadratic Lie algebras include semi-simple Lie algebras, or the semi-direct product $\mathfrak{g} = \mathfrak{s} \ltimes \mathfrak{s}^*$ of a Lie algebra $\mathfrak{s}$ with its dual. Using B we can define the Clifford algebra $\mathrm{Cl}(\mathfrak{g})$. Duflo's factor $J^{1/2}(\xi)$ arises as the Berezin integral of $\exp(q(\lambda(\xi))) \in \mathrm{Cl}(\mathfrak{g})$, where $q : \wedge(\mathfrak{g}) \to \mathrm{Cl}(\mathfrak{g})$ is the quantization map, and $\lambda : \mathfrak{g} \to \wedge^2\mathfrak{g}$ is the map dual to the Lie bracket.

2. Clifford algebras

Let V be a finite-dimensional real vector space, equipped with a non-degenerate symmetric bilinear form B. Fix a basis $e_a \in V$ and let $e^a \in V$ be the dual basis. We denote by $\mathfrak{o}(V) \subset \mathrm{End}(V)$ the space of endomorphisms A of V that are skew-symmetric with respect to B. For any $A \in \mathfrak{o}(V)$ we denote its components by $A_{ab} = B(e_a, Ae_b)$. Consider the function $\mathcal{S} : \mathfrak{o}(V) \to \wedge(V)$ given by

$$\mathcal{S}(A) = \det{}^{1/2}(j(A)) \ \exp_{\wedge(V)}\left(\frac{1}{2} f(A)_{ab} e^a \wedge e^b\right)$$

(using summation convention), where

$$f(z) = (\ln j)'(z) = \frac{1}{2}\coth(\frac{z}{2}) - \frac{1}{z}. \tag{2.1}$$

In turns out that, despite the singularities of the exponential, $\mathcal{S}$ is a global analytic function on all of $\mathfrak{o}(V)$. It has the following nice property. Let $\mathrm{Cl}(V)$ denote the Clifford algebra of V, defined as a quotient of the tensor algebra $T(V)$ by the relations $vv' + v'v = B(v, v')$. The inclusion of $\wedge(V)$ into $T(V)$ as totally anti-symmetric tensors, followed by the quotient map to $\mathrm{Cl}(V)$, gives a vector space isomorphism

$$q : \ \wedge(V) \to \mathrm{Cl}(V)$$

known as the *quantization map*. Then $\mathcal{S}(A)$ relates the exponentials of quadratic elements $1/2 A_{ab} e^a \wedge e^b$ in the exterior algebra with the exponentials of the corresponding elements $1/2 A_{ab} e^a e^b$ in the Clifford algebra:

$$\exp_{\mathrm{Cl}(V)}(1/2 A_{ab} e^a e^b) = q\left(\iota(\mathcal{S}(A)) \exp_{\wedge(V)}(1/2 A_{ab} e^a \wedge e^b)\right). \tag{2.2}$$

Here $\iota : \wedge(V) \to \mathrm{End}(V)$ is the contraction operator. In fact, one may add linear terms to the exponent: Let E be some vector space of "parameters", and $\phi^a \in E$. Then the following identity holds in the $\mathbb{Z}_2$-graded tensor product $\mathrm{Cl}(V) \otimes \wedge(E)$:

$$\exp(1/2A_{ab}e^a e^b + e_a \otimes \phi^a) = q\Big(\iota(\mathcal{S}(A))\exp(1/2A_{ab}e^a \wedge e^b + e_a \otimes \phi^a)\Big).$$

3. Quadratic Lie algebras

Let us now consider the case $V = \mathfrak{g}$ of a quadratic Lie algebra. Invariance of the bilinear form B means that the the adjoint representation $\mathrm{ad} : \mathfrak{g} \to \mathrm{End}(\mathfrak{g})$ takes values in $\mathfrak{o}(\mathfrak{g})$, or equivalently that the structure constants $f_{abc} = B(e_a, [e_b, e_c])$ are invariant under cyclic permutations of the indices a, b, c. We specialize (2.2) to $A = \mathrm{ad}_\xi$ for $\xi \in \mathfrak{g}$, so that $\lambda^{\mathfrak{g}} : \mathfrak{g} \to \wedge^2\mathfrak{g}$,

$$\lambda^{\mathfrak{g}}(\xi) = 1/2(\mathrm{ad}_\xi)_{ab}e^a \wedge e^b$$

is the map dual to the Lie bracket. Also, take $E = T^*_\xi\mathfrak{g}$ and $\phi^a = -d\xi^a$, where $\xi^a = B(\xi, e^a)$ are the coordinate functions. Then our formula become the following identity in $\mathrm{Cl}(\mathfrak{g}) \otimes \Omega(\mathfrak{g})$

$$\exp\left(q(\lambda^{\mathfrak{g}}) - e_a\, d\xi^a\right) = q\Big(\iota(\mathcal{S}^{\mathfrak{g}})\exp(\lambda^{\mathfrak{g}} - e_a\, d\xi^a)\Big), \tag{3.1}$$

where $\mathcal{S}^{\mathfrak{g}} = \mathcal{S} \circ \mathrm{ad} : \mathfrak{g} \to \wedge\mathfrak{g}$. Consider now the following cubic element in the Clifford algebra,

$$\mathcal{C} = \frac{1}{6}f_{abc}e^a\, e^b\, e^c \in \mathrm{Cl}(\mathfrak{g}).$$

A beautiful observation of Kostant-Sternberg [8] says that $\mathcal{C}$ squares to a *constant*:

$$\mathcal{C}^2 = -\frac{1}{48}f_{abc}f^{abc}.$$

It follows that the graded commutator $\mathrm{d}^{\mathrm{Cl}\,\mathfrak{g}} := [\mathcal{C}, \cdot]$ defines a differential on $\mathrm{Cl}(\mathfrak{g})$. This *Clifford differential* is compatible with the filtration of $\mathrm{Cl}(\mathfrak{g})$, and the induced differential $\mathrm{d}^{\wedge\mathfrak{g}}$ on the associated graded algebra $\mathrm{gr}(\mathrm{Cl}(\mathfrak{g})) = \wedge\mathfrak{g}$ is nothing but the Lie algebra differential. Let d^{Rh} denote the exterior differential on the deRham complex $\Omega(\mathfrak{g})$.

It is easily verified that $\lambda^{\mathfrak{g}} - e_a\, d\xi^a \in \wedge\mathfrak{g} \otimes \Omega(\mathfrak{g})$ is closed for the differential $\mathrm{d}^{\wedge\mathfrak{g}} + \mathrm{d}^{\mathrm{Rh}}$, while $q(\lambda^{\mathfrak{g}}) - e_a\, d\xi^a \in \mathrm{Cl}(\mathfrak{g}) \otimes \Omega(\mathfrak{g})$ is closed under $\mathrm{d}^{\mathrm{Cl}(\mathfrak{g})} + \mathrm{d}^{\mathrm{Rh}}$. Together with (3.1), this leads to a number of consistency conditions for the function $\mathcal{S}^{\mathfrak{g}}$. One of these conditions gives a solution of the *classical dynamical Yang-Baxter equation* (CDYBE): Let $\mathfrak{r} : \mathfrak{g} \to \mathfrak{o}(\mathfrak{g})$ be the meromorphic function $\mathfrak{r}^{\mathfrak{g}}(\xi) = f(\mathrm{ad}_\xi)$ appearing in the exponential factor of $\mathcal{S}^{\mathfrak{g}}$. Then

$$\mathrm{cycl}_{abc}\left(\frac{\partial \mathfrak{r}_{ab}}{\partial \xi^c} - \mathfrak{r}_{ak}f^{kl}_b\mathfrak{r}_{lc}\right) = -\frac{1}{4}f_{abc} \tag{3.2}$$

where cycl_{abc} denotes the sum over cyclic permutations of a, b, c. This solution of the CDYBE was obtained by Etingof-Varchenko [5] and in [1] by different methods.

In Etingof-Schiffmann [6], it is shown that $\mathfrak{r}_{ab}$ is in fact the *unique* solution of this particular CDYBE, up to gauge transformation. More general CDYBE's are associated to a pair $\mathfrak{h} \subset \mathfrak{g}$ of Lie algebras, here $\mathfrak{h} = \mathfrak{g}$. The proof sketched above can be modified to produce some of these more general solutions.

4. The non-commutative Weil algebra

Using B to identify the Lie algebra $\mathfrak{g}$ with its dual $\mathfrak{g}^*$, the Weil algebra of $\mathfrak{g}$ is the $\mathbb{Z}$-graded $\mathfrak{g}$-module given as a tensor product

$$W\mathfrak{g} = S\mathfrak{g} \otimes \wedge\mathfrak{g},$$

where generators of $S\mathfrak{g}$ are assigned degree 2. Let L_ξ^W for $\xi \in \mathfrak{g}$ denote the generators for the $\mathfrak{g}$-action on $W\mathfrak{g}$, and $\iota_\xi^W = 1 \otimes \iota_\xi$ the contraction operators. The *Weil differential* d^W is a derivation of degree 1, uniquely characterized by its properties $\mathrm{d}^W \circ \mathrm{d}^W = 0$ and $\mathrm{d}^W(1 \otimes \xi) = \xi \otimes 1$ for $\xi \in \mathfrak{g}$. The Weil algebra $W\mathfrak{g}$ with these three types of derivations is an example of a $\mathfrak{g}$-differential algebra: That is, $L_\xi^W, \iota_\xi^W, \mathrm{d}^W$ satisfy relations similar to contraction operators, Lie derivatives, and de Rham differential for a manifold with group action.

In [1], we introduced the following non-commutative version of the Weil algebra,

$$\mathcal{W}\mathfrak{g} = U\mathfrak{g} \otimes \mathrm{Cl}(\mathfrak{g}).$$

It carries a $\mathbb{Z}$-filtration, where generators of $U(\mathfrak{g})$ are assigned filtration degree 2, with associated graded algebra $\mathrm{gr}(\mathcal{W}\mathfrak{g}) = W\mathfrak{g}$. Moreover, it carries a $\mathbb{Z}_2$-grading, compatible with the $\mathbb{Z}$-filtration in the sense of [8]. Define contraction operators as $\mathbb{Z}_2$-graded commutators $\iota_\xi^{\mathcal{W}} = [1 \otimes \xi, \cdot\,]$, let $L_\xi^{\mathcal{W}}$ be the generators for the natural $\mathfrak{g}$-module structure, and set $\mathrm{d}^{\mathcal{W}} = [\mathcal{D}, \cdot\,]$ where

$$\mathcal{D} = e_a \otimes e^a - 1 \otimes \mathcal{C} \in \mathcal{W}\mathfrak{g}$$

is the *cubic Dirac operator* [7]. Its square

$$\mathcal{D}^2 = \frac{1}{2} e_a e^a \otimes 1 - \frac{1}{48} f_{abc} f^{abc}$$

is in the center of $\mathcal{W}\mathfrak{g}$, hence $\mathrm{d}^{\mathcal{W}}$ is a differential. As it turns out, $\mathcal{W}\mathfrak{g}$ is again a $\mathfrak{g}$-differential algebra. The derivations $\mathrm{d}^{\mathcal{W}}, \iota_\xi^{\mathcal{W}}, L_\xi^{\mathcal{W}}$ respect the $\mathbb{Z}$-filtration, and the induced derivations on the associated graded algebra are just the standard derivations for the Weil algebra $W\mathfrak{g}$.

The vector space isomorphism $\mathrm{sym} \otimes q : W\mathfrak{g} \to \mathcal{W}\mathfrak{g}$ intertwines the contraction operators and Lie derivatives, but not the differentials. There does exist, however, a better "quantization map" $\mathcal{Q} : W\mathfrak{g} \to \mathcal{W}\mathfrak{g}$ that is also a chain map. Using our function $\mathcal{S}^{\mathfrak{g}} \in C^\infty(\mathfrak{g}) \otimes \wedge\mathfrak{g}$, let $\iota(\widehat{\mathcal{S}^{\mathfrak{g}}})$ denote the operator on $W\mathfrak{g}$, where the $\wedge\mathfrak{g}$-factor acts by contraction on $\wedge\mathfrak{g}$ and the $C^\infty(\mathfrak{g})$-factor as an infinite order differential operator.

Theorem. [1] *The quantization map*

$$\mathcal{Q} := (\mathrm{sym} \otimes q) \circ \iota(\widehat{\mathcal{S}^{\mathfrak{g}}}) : W\mathfrak{g} \to \mathcal{W}\mathfrak{g}$$

intertwines the contraction operators, Lie derivatives, and differentials on $W\mathfrak{g}$ *and on* $\mathcal{W}\mathfrak{g}$.

The fact that $\mathcal{Q}$ intertwines the two differentials d^W, $\mathrm{d}^{\mathcal{W}}$ relies on a number of special properties of the function $\mathcal{S}^{\mathfrak{g}}$, including the CDYBE.

Put differently, the quantization map $\mathcal{Q}$ defines a new, graded non-commutative ring structure on the Weil algebra $W\mathfrak{g}$, in such a way that the derivations $\iota_\xi^W, L_\xi^W, \mathrm{d}^W$ are still derivations for the new ring structure, and in fact become *inner* derivations. Notice that $\mathcal{Q}$ restricts to the quantization map for Clifford algebras $q : \wedge\mathfrak{g} \to \mathrm{Cl}(\mathfrak{g})$ on the second factor and to the Duflo map on the first factor, but is not just the product of these two maps.

5. Equivariant cohomology

H. Cartan in [3] introduced the Weil algebra $W\mathfrak{g}$ as an algebraic model for the algebra of differential forms on the classifying bundle EG, at least in the case G compact.

In particular, it can be used to compute the equivariant cohomology $H_G(M)$ (with real coefficients) for any G-manifold M. Let $\iota_\xi^{Rh}, L_\xi^{Rh}, \mathrm{d}^{Rh}$ denote the contraction operators, Lie derivatives, and differential on the de Rham complex $\Omega(M)$ of differential forms. Let

$$H_{\mathfrak{g}}(M) = H((W\mathfrak{g} \otimes \Omega(M))_{\mathrm{basic}}, \mathrm{d}^W + \mathrm{d}^{Rh})$$

where $(W\mathfrak{g} \otimes \Omega(M))_{\mathrm{basic}}$ is the subspace annihilated by all Lie derivatives $L_\xi^W + L_\xi^{Rh}$ and all contraction operators $\iota_\xi^W + \iota_\xi^{Rh}$. Cartan's result says that $H_{\mathfrak{g}}(M) = H_G(M, \mathbb{R})$ provided G is compact.

More generally, we can define $H_{\mathfrak{g}}(\mathcal{A})$ for any $\mathfrak{g}$-differential algebra $\mathcal{A}$. Let $\mathcal{H}_{\mathfrak{g}}(\mathcal{A})$ be defined by replacing $W\mathfrak{g}$ with $\mathcal{W}\mathfrak{g}$. The quantization map $\mathcal{Q} : W\mathfrak{g} \to \mathcal{W}\mathfrak{g}$ induces a map $\mathcal{Q} : H_{\mathfrak{g}}(\mathcal{A}) \to \mathcal{H}_{\mathfrak{g}}(\mathcal{A})$.

Theorem. [1] *For any* $\mathfrak{g}$*-differential algebra* $\mathcal{A}$*, the vector space isomorphism* $\mathcal{Q} : H_{\mathfrak{g}}(\mathcal{A}) \to \mathcal{H}_{\mathfrak{g}}(\mathcal{A})$ *is in fact an algebra isomorphism.*

Our proof is by construction of an explicit chain homotopy between the two maps $W\mathfrak{g} \otimes W\mathfrak{g} \to \mathcal{W}\mathfrak{g}$ given by "quantization followed by multiplication" and "multiplication followed by quantization", respectively. Taking $\mathcal{A}$ to be the trivial $\mathfrak{g}$-differential algebra (i.e. $\mathcal{A} = \Omega(\mathrm{point})$), the statement specializes to Duflo's theorem for quadratic $\mathfrak{g}$.

References

[1] A. Alekseev & E. Meinrenken, The non-commutative Weil algebra, *Invent. Math.* 139 (2000), 135–172.

[2] A. Alekseev & E. Meinrenken, Clifford algebras and the classical dynamical Yang-Baxter equation (in preparation).
[3] H. Cartan, Notions d'algèbre différentielle; application aux groupes de Lie et aux variétés où opère un groupe de Lie., Colloque de topologie (espaces fibrés), Bruxelles, (1950).
[4] M. Duflo, Opérateurs différentiels bi-invariants sur un groupe de Lie, Ann. Sci. École Norm. Sup. 10 (1977), 265–288.
[5] P. Etingof & A. Varchenko, Geometry and classification of solutions of the classical dynamical Yang-Baxter equation, *Comm. Math. Phys.*, 192 (1988), 77-120.
[6] P. Etingof & O. Schiffmann, On the moduli space of classical dynamical r-matrices, *Math. Res. Lett.* 8 (2001), 157–170.
[7] B. Kostant, A cubic Dirac operator and the emergence of Euler number multiplets of representations for equal rank subgroups, *Duke Math. J.* 100 (1999), no. 3, 447–501.
[8] B. Kostant & S. Sternberg, Symplectic reduction, BRS cohomology, and infinite-dimensional Clifford algebras, *Ann. Physics* 176 (1987), no. 1, 49–113.

ICM 2002 · Vol. II · 643–654

Representations of Yangians Associated with Skew Young Diagrams

Maxim Nazarov*

Abstract

The Yangian of the Lie algebra $\mathfrak{gl}_N$ has a distinguished family of irreducible finite-dimensional representations, called elementary representations. They are parametrized by pairs, consisting of a skew Young diagram and a complex number. Each of these representations has an explicit realization, it extends the classical realization of the irreducible polynomial representations of $\mathfrak{gl}_N$ by means of the Young symmetrizers. We explicitly construct analogues of these elementary representations for the twisted Yangian, which corresponds to the Lie algebra $\mathfrak{so}_N$. Our construction provides solutions to several open problems in the classical representation theory. In particular, we obtain analogues of the Young symmetrizers for the Brauer centralizer algebra.

2000 Mathematics Subject Classification: 17B35, 17B37, 20C30, 22E46.
Keywords and Phrases: Branching rules, Brauer algebra, Classical groups, Intertwining operators, Reflection equation, Yangians, Young symmetrizers.

1. Yangian of the general linear Lie algebra

1.1. For each simple finite-dimensional Lie algebra $\mathfrak{g}$ over the field $\mathbb{C}$, Drinfeld [4] introduced a canonical deformation of the universal enveloping algebra of the polynomial current Lie algebra $\mathfrak{g}[x]$. This deformation is a certain Hopf algebra over $\mathbb{C}$, denoted by $\mathrm{Y}(\mathfrak{g})$ and called the *Yangian* of the simple Lie algebra $\mathfrak{g}$. Now consider the general linear Lie algebra $\mathfrak{gl}_N$, it contains the special linear Lie algebra $\mathfrak{sl}_N$ as a subalgebra. The Hopf algebra which is called the Yangian of the reductive Lie algebra $\mathfrak{gl}_N$ and is denoted by $\mathrm{Y}(\mathfrak{gl}_N)$, was considered in the earlier works of mathematical physicists from St.-Petersburg, see for instance [6]. The Hopf algebra $\mathrm{Y}(\mathfrak{gl}_N)$ is a deformation of the universal enveloping algebra of the Lie algebra $\mathfrak{gl}_N[x]$, and the Yangian $\mathrm{Y}(\mathfrak{sl}_N)$ of the simple Lie algebra $\mathfrak{sl}_N$ is a Hopf subalgebra of $\mathrm{Y}(\mathfrak{gl}_N)$. Throughout this article, we assume that N is a positive integer.

*Department of Mathematics, University of York, York YO10 5DD, England. E-mail: mln1@york.ac.uk

The unital associative algebra $\mathrm{Y}(\mathfrak{gl}_N)$ over $\mathbb{C}$ has a family of generators $T_{ij}^{(a)}$ where $a = 1, 2, \ldots$ and $i, j = 1, \ldots, N$. The defining relations for these generators can be written in terms of the formal power series

$$T_{ij}(x) = \delta_{ij} \cdot 1 + T_{ij}^{(1)} x^{-1} + T_{ij}^{(2)} x^{-2} + \ldots \in \mathrm{Y}(\mathfrak{gl}_N)\,[[x^{-1}]]\,. \tag{1.1}$$

Here x is the formal parameter. Let y be another formal parameter, then the defining relations in the associative algebra $\mathrm{Y}(\mathfrak{gl}_N)$ can be written as

$$(x-y) \cdot [T_{ij}(x), T_{kl}(y)] = T_{kj}(x) T_{il}(y) - T_{kj}(y) T_{il}(x)\,, \tag{1.2}$$

where $i, j, k, l = 1, \ldots, N$. The square brackets in (1.2) denote usual commutator. In terms of the formal series (1.1), the coproduct $\Delta : \mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{Y}(\mathfrak{gl}_N) \otimes \mathrm{Y}(\mathfrak{gl}_N)$ is defined by

$$\Delta\big(T_{ij}(x)\big) = \sum_{k=1}^{N} T_{ik}(x) \otimes T_{kj}(x)\,; \tag{1.3}$$

the tensor product on the right hand side of the equality (1.3) is taken over the subalgebra $\mathbb{C}[[x^{-1}]] \subset \mathrm{Y}(\mathfrak{gl}_N)\,[[x^{-1}]]$. The counit homomorphism $\varepsilon : \mathrm{Y}(\mathfrak{gl}_N) \to \mathbb{C}$ is determined by the assignment $\varepsilon : T_{ij}(u) \mapsto \delta_{ij} \cdot 1$.

For each i and j one can determine a formal power series $\widetilde{T}_{ij}(x)$ in x^{-1} with the coefficients in $\mathrm{Y}(\mathfrak{gl}_N)$ and the leading term δ_{ij}, by the system of equations

$$\sum_{k=1}^{N} T_{ik}(x)\, \widetilde{T}_{kj}(x) = \delta_{ij} \quad \text{where} \quad i, j = 1, \ldots, N.$$

The antipode S on $\mathrm{Y}(\mathfrak{gl}_N)$ is the anti-automorphism of the algebra $\mathrm{Y}(\mathfrak{gl}_N)$, defined by the assignment $\mathrm{S} : T_{ij}(x) \mapsto \widetilde{T}_{ij}(x)$. We also use the involutive automorphism ξ_N of the algebra $\mathrm{Y}(\mathfrak{gl}_N)$, defined by the assignment $\xi_N : T_{ij}(x) \mapsto \widetilde{T}_{ij}(-x)$.

Take any formal power series $f(x) \in \mathbb{C}[[x^{-1}]]$ with the leading term 1. The assignment

$$T_{ij}(x) \mapsto f(x) \cdot T_{ij}(x) \tag{1.4}$$

defines an automorphism of the algebra $\mathrm{Y}(\mathfrak{gl}_N)$, this follows from (1.1) and (1.2). The Yangian $\mathrm{Y}(\mathfrak{sl}_N)$ is the subalgebra in $\mathrm{Y}(\mathfrak{gl}_N)$ consisting of all elements, which are invariant under every automorphism (1.4).

It also follows from (1.1) and (1.2) that for any $z \in \mathbb{C}$, the assignment

$$\tau_z : T_{ij}(x) \mapsto T_{ij}(x - z)$$

defines an automorphism τ_z of the algebra $\mathrm{Y}(\mathfrak{gl}_N)$. Here the formal power series in $(x-z)^{-1}$ should be re-expanded in x^{-1}. Regard the matrix units $E_{ij} \in \mathfrak{gl}_N$ as generators of the universal enveloping algebra $\mathrm{U}(\mathfrak{gl}_N)$. The assignment

$$\alpha_N : T_{ij}(x) \mapsto \delta_{ij} \cdot 1 - E_{ji}\, x^{-1}$$

defines a homomorphism $\alpha_N : \mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{U}(\mathfrak{gl}_N)$. By definition, the homomorphism α_N is surjective. For more details and references on the definition of the Yangian $\mathrm{Y}(\mathfrak{gl}_N)$, see [7].

1.2. Let $\nu = (\nu_1, \nu_2, \dots)$ be any partition. As usual, the parts of ν are arranged in the non-increasing order: $\nu_1 \geqslant \nu_2 \geqslant \dots \geqslant 0$. Let $\nu' = (\nu'_1, \nu'_2, \dots)$ be the partition conjugate to ν. In particular, ν'_1 is the number of non-zero parts of the partition ν. An irreducible module over the Lie algebra $\mathfrak{gl}_N$ is called *polynomial*, if it is equivalent to a submodule in the tensor product of n copies of the defining $\mathfrak{gl}_N$-module $\mathbb{C}^N$, for some integer $n \geqslant 0$. The irreducible polynomial $\mathfrak{gl}_N$-modules are parametrized by partitions ν such that $\nu'_1 \leqslant N$. Here $n = \nu_1 + \nu_2 + \dots$. Let V_ν be the irreducible module corresponding to ν. This $\mathfrak{gl}_N$-module is of highest weight $(\nu_1, \dots, \nu_N)$. Here we choose the Borel subalgebra in $\mathfrak{gl}_N$ consisting of the upper triangular matrices, and fix the basis of the diagonal matrix units $E_{11}, \dots, E_{NN}$ in the corresponding Cartan subalgebra of $\mathfrak{gl}_N$.

Take any non-negative integer M. Let the indices i and j range over the set $\{1, \dots, N+M\}$. Fix the basis of the matrix units E_{ij} in the Lie algebra $\mathfrak{gl}_{N+M}$. We suppose that the subalgebras $\mathfrak{gl}_N$ and $\mathfrak{gl}_M$ in $\mathfrak{gl}_{N+M}$ are spanned by elements E_{ij} where respectively $i,j = 1, \dots, N$ and $i,j = N+1, \dots, N+M$. Let λ and μ be two partitions, such that $\lambda'_1 \leqslant N+M$ and $\mu'_1 \leqslant M$. Consider the irreducible modules V_λ and V_μ over the Lie algebras $\mathfrak{gl}_{N+M}$ and $\mathfrak{gl}_M$. The vector space

$$\mathrm{Hom}_{\mathfrak{gl}_M}(V_\mu, V_\lambda) \tag{1.5}$$

comes with a natural action of the Lie algebra $\mathfrak{gl}_N$. This action of $\mathfrak{gl}_N$ may be reducible. The vector space (1.5) is non-zero, if and only if $\lambda_k \geqslant \mu_k$ and $\lambda'_k - \mu'_k \leqslant N$ for each $k = 1, 2, \dots$; see for instance [8].

Denote by $\mathrm{A}_N(M)$ the centralizer of the subalgebra $\mathrm{U}(\mathfrak{gl}_M) \subset \mathrm{U}(\mathfrak{gl}_{N+M})$. The centralizer $\mathrm{A}_N(M) \subset \mathrm{U}(\mathfrak{gl}_{N+M})$ contains $\mathrm{U}(\mathfrak{gl}_N)$ as a subalgebra, and acts naturally in the vector space (1.5). This action is irreducible. For every M, Olshanski [16] defined a homomorphism of associative algebras $\mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{A}_N(M)$. Along with the centre of the algebra $\mathrm{U}(\mathfrak{gl}_{N+M})$, the image of this homomorphism generates the algebra $\mathrm{A}_N(M)$. We use a version of this homomorphism, it is denoted by α_{NM}.

The subalgebra in $\mathrm{Y}(\mathfrak{gl}_{N+M})$ generated by $T_{ij}^{(a)}$ where $i,j = 1, \dots, N$, by definition coincides with the Yangian $\mathrm{Y}(\mathfrak{gl}_N)$. Denote by φ_M this natural embedding $\mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{Y}(\mathfrak{gl}_{N+M})$. Consider also the involutive automorphism ξ_{N+M} of the algebra $\mathrm{Y}(\mathfrak{gl}_{N+M})$. The image of the homomorphism

$$\alpha_{N+M} \circ \xi_{N+M} \circ \varphi_M : \mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{U}(\mathfrak{gl}_{N+M})$$

belongs to the subalgebra $\mathrm{A}_N(M) \subset \mathrm{U}(\mathfrak{gl}_{N+M})$. Moreover, this image along with the centre of the algebra $\mathrm{U}(\mathfrak{gl}_{N+M})$, generates the subalgebra $\mathrm{A}_N(M)$. For the proofs of these claims, see [9]. We use the homomorphism $\mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{U}(\mathfrak{gl}_{N+M})$

$$\alpha_{NM} = \alpha_{N+M} \circ \xi_{N+M} \circ \varphi_M \circ \xi_N . \tag{1.6}$$

When $M = 0$, the homomorphism (1.6) coincides with α_N. The intersection of the kernels of all homomorphisms $\alpha_{N0}, \alpha_{N1}, \alpha_{N2}, \dots$ is zero [16].

1.3. The $\mathrm{A}_N(M)$-module (1.5) depends on the partitions λ and μ via the *skew Young diagram*

$$\omega = \{\, (i,j) \in \mathbb{Z}^2 \mid i \geqslant 1,\ \lambda_i \geqslant j > \mu_i \,\} .$$

When $\mu = (0,0,\ldots)$, this is the usual Young diagram of the partition λ. Consider the $\mathrm{Y}(\mathfrak{gl}_N)$-module obtained from the $\mathrm{A}_N(M)$-module (1.5) by pulling back through the homomorphism $\alpha_{NM} \circ \tau_z : \mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{A}_N(M)$. Since the central elements of $\mathrm{U}(\mathfrak{gl}_{N+M})$ act in (1.5) as scalar operators, this $\mathrm{Y}(\mathfrak{gl}_N)$-module is irreducible. It is denoted by $V_\omega(z)$, and is called an *elementary module.* Its equivalence class does not depend on the choice of the integer M, such that $\lambda'_1 \leqslant N + M$ and $\mu'_1 \leqslant M$.

The elementary modules are distinguished amongst all irreducible $\mathrm{Y}(\mathfrak{gl}_N)$-modules by the following theorem. Consider the chain of algebras

$$\mathrm{Y}(\mathfrak{gl}_1) \subset \mathrm{Y}(\mathfrak{gl}_2) \subset \ldots \subset \mathrm{Y}(\mathfrak{gl}_N)\,. \tag{1.7}$$

Here for every $k = 1, \ldots, N-1$ we use the embedding $\varphi_1 : \mathrm{Y}(\mathfrak{gl}_k) \to \mathrm{Y}(\mathfrak{gl}_{k+1})$. Consider the subalgebra of $\mathrm{Y}(\mathfrak{gl}_N)$ generated by the centres of all algebras in the chain (1.7), it is called the *Gelfand-Zetlin subalgebra.* This subalgebra is maximal commutative in $\mathrm{Y}(\mathfrak{gl}_N)$; see [3] and [13]. Take any finite-dimensional module W over the Yangian $\mathrm{Y}(\mathfrak{gl}_N)$.

Theorem 1. *Two conditions on the* $\mathrm{Y}(\mathfrak{gl}_N)$*-module* W *are equivalent:*

a) W *is irreducible, and the action of the Gelfand-Zetlin subalgebra of* $\mathrm{Y}(\mathfrak{gl}_N)$ *in* W *is semi-simple;*

b) W *is obtained by pulling back through some automorphism* (1.4) *from the tensor product*

$$V_{\omega_1}(z_1) \otimes \ldots \otimes V_{\omega_m}(z_m) \tag{1.8}$$

of elementary $\mathrm{Y}(\mathfrak{gl}_N)$*-modules, for some skew Young diagrams* $\omega_1, \ldots, \omega_m$ *and for some complex numbers* $z_1, \ldots, z_m$ *such that* $z_k - z_l \notin \mathbb{Z}$ *for all* $k \neq l$.

This characterization of irreducible finite-dimensional $\mathrm{Y}(\mathfrak{gl}_N)$-modules with semi-simple action of the Gelfand-Zetlin subalgebra was conjectured by Cherednik, and was proved by him [3] under certain extra conditions on the module W. In full generality, Theorem 1 was proved in [14]. An irreducibility criterion for the $\mathrm{Y}(\mathfrak{gl}_N)$-module (1.8) with arbitrary parameters $z_1, \ldots, z_m$ was given in [15].

The classification of all irreducible finite-dimensional $\mathrm{Y}(\mathfrak{gl}_N)$-modules has been given by Drinfeld [5]. However, the general structure of these modules needs a better understanding. For instance, the dimensions of these modules are not explicitly known in general. The tensor products (1.8) provide a wide class of irreducible $\mathrm{Y}(\mathfrak{gl}_N)$-modules, which can be constructed explicitly.

1.4. The $\mathrm{Y}(\mathfrak{gl}_N)$-module $V_\omega(z)$ has an explicit realization. It extends the classical realization of irreducible $\mathfrak{gl}_N$-module V_ν by means of the Young symmetrizers [20].

Let us use the standard graphic representation of Young diagrams on the plane $\mathbb{R}^2$ with two matrix style coordinates. The first coordinate increases from top to bottom, the second coordinate increases from left to right. The element $(i,j) \in \omega$ is represented by the unit box with the bottom right corner at the point $(i,j) \in \mathbb{R}^2$.

Suppose the set ω consists of n elements. Consider the *column tableau* of shape ω. It is obtained by filling the boxes of ω with numbers $1, \ldots, n$ consecutively by columns from left to right, downwards in every column. Denote this tableau by Ω.

For each $k = 1, \ldots, n$ put $c_k = j - i$ if the box $(i, j) \in \omega$ is filled with the number k in the tableau Ω. The difference $j - i$ is called the *content* of the box (i, j) of the diagram ω. Our choice of the tableau Ω provides an ordering of the collection of all contents of ω. In the above figure, on the left we show the column tableau Ω for the partitions $\lambda = (5,3,3,3,3,0,0,\ldots)$ and $\mu = (3,3,2,0,0,\ldots)$. On the right we indicate the contents of all boxes of ω.

Introduce n complex variables $t_1, \ldots, t_n$ with the constraints $t_k = t_l$ for all k and l occuring in the same column of Ω. The number of independent variables among $t_1, \ldots, t_n$ equals the number of non-empty columns in the diagram ω. Order lexicographically the set of all pairs (k,l) with $1 \leqslant k < l \leqslant n$. Take the ordered product over this set,

$$\prod_{1 \leqslant k < l \leqslant n}^{\rightarrow} \left(1 - \frac{P_{kl}}{c_k - c_l + t_k - t_l}\right) \tag{1.9}$$

where P_{kl} denotes the operator in the space $(\mathbb{C}^N)^{\otimes n}$ exchanging the kth and lth tensor factors. Consider (1.9) as a function of the constrained variables $t_1, \ldots, t_n$.

Proposition 1. *The rational function* (1.9) *is regular at* $t_1 = \ldots = t_n$.

The rational function (1.9) depends only on the differences $t_k - t_l$. Denote the value of (1.9) at $t_1 = \ldots = t_n$ by E_Ω. Note that for any λ and μ, the linear operator E_Ω in the vector space $(\mathbb{C}^N)^{\otimes n}$ does not depend on M. For the proof of Proposition 1, see [15]. It provides an explicit expression for the operator E_Ω.

Suppose that $\mu = (0,0,\ldots)$. In this special case, there is another expression for the operator E_Ω. Consider the action of the symmetric group S_n on $(\mathbb{C}^N)^{\otimes n}$ by permutations of the tensor factors. For any $s \in S_n$, denote by P_s the corresponding operator in $(\mathbb{C}^N)^{\otimes n}$. Let S_λ (respectively S'_λ) be the subgroup in S_n preserving, as sets, the collections of numbers appearing in every row (every column) of the tableau Ω. Put

$$X_\lambda = \sum_{s \in S_\lambda} P_s \quad \text{and} \quad Y_\lambda = \sum_{s \in S'_\lambda} P_s \cdot \operatorname{sgn} s$$

where $\operatorname{sgn} s = \pm 1$ is the sign of the permutation s. The product $X_\lambda Y_\lambda$ is the *Young symmetrizer* in $(\mathbb{C}^N)^{\otimes n}$ corresponding to the tableau Ω. We have the equality

$$E_\Omega = Y_\lambda X_\lambda Y_\lambda \,/\, \lambda_1'!\, \lambda_2'! \ldots, \tag{1.10}$$

see [10]. In this case, the image of the operator E_Ω in $(\mathbb{C}^N)^{\otimes n}$ is equivalent to V_λ as $\mathfrak{gl}_N$-module, see [20]. Here the action of the Lie algebra $\mathfrak{gl}_N$ in $(\mathbb{C}^N)^{\otimes n}$ is standard.

1.5. By pulling the standard action of $\mathrm{U}(\mathfrak{gl}_N)$ in the space $\mathbb{C}^N$ back through the homomorphism

$$\alpha_N \circ \tau_z : \mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{U}(\mathfrak{gl}_N),$$

we obtain a module over the algebra $\mathrm{Y}(\mathfrak{gl}_N)$, which is denoted by $V(z)$ and called an *evaluation module*. We have $V(z) = V_\omega(z)$ for $\lambda = (1,0,\dots)$ and $\mu = (0,0,\dots)$. For any partitions λ and μ, the operator E_Ω has the following interpretation, in terms of the tensor products of evaluation modules over the Hopf algebra $\mathrm{Y}(\mathfrak{gl}_N)$. Let P_0 be the operator in $(\mathbb{C}^N)^{\otimes n}$ reversing the order of the tensor factors.

Proposition 2. *The operator $E_\Omega P_0$ is an intertwiner of the $\mathrm{Y}(\mathfrak{gl}_N)$-modules*

$$V(c_n + z) \otimes \dots \otimes V(c_1 + z) \;\longrightarrow\; V(c_1 + z) \otimes \dots \otimes V(c_n + z)\,.$$

By Proposition 2, the image of the operator E_Ω is a submodule in the tensor product of evaluation $\mathrm{Y}(\mathfrak{gl}_N)$-modules $V(c_1 + z) \otimes \dots \otimes V(c_n + z)$. Denote this $\mathrm{Y}(\mathfrak{gl}_N)$-submodule by $V_\Omega(z)$. For any λ and μ, we have the following theorem. Put

$$f_\mu(x) \;=\; \prod_{k \geqslant 1} \frac{(x - \mu_k + k)(x + k - 1)}{(x - \mu_k + k - 1)(x + k)}\,. \tag{1.11}$$

This rational function of x expands as a power series in x^{-1} with the leading term 1.

Theorem 2. *The $\mathrm{Y}(\mathfrak{gl}_N)$-module $V_\Omega(z)$ is equivalent to the elementary module $V_\omega(z)$, pulled back through the automorphism of the algebra $\mathrm{Y}(\mathfrak{gl}_N)$ defined by* (1.4)*, where $f(x) = f_\mu(x - z)$.*

This theorem is due to Cherednik [3], see also [12]. It provides an explicit realization of the elementary $\mathrm{Y}(\mathfrak{gl}_N)$-module $V_\omega(z)$ as a subspace in $(\mathbb{C}^N)^{\otimes n}$. It also shows that the $\mathrm{Y}(\mathfrak{gl}_N)$-module $V_\Omega(z)$ is irreducible, cf. [15]. The isomorphism between the $\mathrm{Y}(\mathfrak{gl}_N)$-module $V_\Omega(z)$, and the pull-back of the $\mathrm{Y}(\mathfrak{gl}_N)$-module $V_\omega(z)$ as in Theorem 1, is unique up to a scalar multiplier.

In Section 2 we give an analogue of Theorem 2 for the orthogonal Lie algebra $\mathfrak{so}_N$, instead of $\mathfrak{gl}_N$. The case of the symplectic Lie algebra $\mathfrak{sp}_N$ is similar to that of $\mathfrak{so}_N$, and is considered in the detailed version [12] of the present article.

For any simple Lie algebra $\mathfrak{g}$ the Yangian $\mathrm{Y}(\mathfrak{g})$ as defined in [4], contains the universal enveloping algebra $\mathrm{U}(\mathfrak{g})$ as a subalgebra. An embedding $\mathrm{U}(\mathfrak{gl}_N) \to \mathrm{Y}(\mathfrak{gl}_N)$ can be defined by

$$E_{ij} \;\mapsto\; -T^{(1)}_{ji}. \tag{1.12}$$

The image of $\mathrm{U}(\mathfrak{sl}_N) \subset \mathrm{U}(\mathfrak{gl}_N)$ under this emdedding belongs to $\mathrm{Y}(\mathfrak{sl}_N) \subset \mathrm{Y}(\mathfrak{gl}_N)$. The homomorphism $\alpha_N : \mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{U}(\mathfrak{gl}_N)$ is identical on the subalgebra $\mathrm{U}(\mathfrak{gl}_N)$. The restriction of α_N to $\mathrm{Y}(\mathfrak{sl}_N)$ provides a homomorphism $\mathrm{Y}(\mathfrak{sl}_N) \to \mathrm{U}(\mathfrak{sl}_N)$, which is identical on the subalgebra $\mathrm{U}(\mathfrak{sl}_N)$. For $\mathfrak{g} \neq \mathfrak{sl}_N$ a homomorphism $\mathrm{Y}(\mathfrak{g}) \to \mathrm{U}(\mathfrak{g})$ identical on the subalgebra $\mathrm{U}(\mathfrak{g}) \subset \mathrm{Y}(\mathfrak{g})$, does not exist [4]. For this reason, instead of the Yangian $\mathrm{Y}(\mathfrak{so}_N)$ from [4], we will consider the *twisted Yangian* $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ from [17]. Here σ is the involutive automorphism of the Lie algebra $\mathfrak{gl}_N$, such that $-\sigma$ is the matrix transposition. Then $\mathfrak{so}_N$ is the subalgebra of σ-fixed points in $\mathfrak{gl}_N$.

2. Twisted Yangian of the orthogonal Lie algebra

2.1. The associative algebra $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ is a deformation of the universal enveloping algebra of the *twisted polynomial current Lie algebra*

$$\{A(x) \in \mathfrak{gl}_N[x] : \sigma(A(x)) = A(-x)\}\,.$$

The deformation $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ is not a Hopf algebra, but a coideal subalgebra in the Hopf algebra $\mathrm{Y}(\mathfrak{gl}_N)$. The definition of the twisted Yangian $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ was motivated by the works of Cherednik [2] and Sklyanin [19] on quantum integrable systems with boundary conditions. This definition was given by Olshanski in [17].

As in Subsection 1.1, let the indices i and j range over the set $\{1, \ldots, N\}$. By definition, $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ is the subalgebra in $\mathrm{Y}(\mathfrak{gl}_N)$ generated by the coefficients of all formal power series

$$\sum_{k=1}^{N} T_{ki}(-x)\, T_{kj}(x) \tag{2.1}$$

in x^{-1}. Due to (1.3), the subalgebra $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ in $\mathrm{Y}(\mathfrak{gl}_N)$ is a right coideal:

$$\Delta\big(\mathrm{Y}(\mathfrak{gl}_N, \sigma)\big) \subset \mathrm{Y}(\mathfrak{gl}_N, \sigma) \otimes \mathrm{Y}(\mathfrak{gl}_N)\,.$$

To give the defining relations for the generators of $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$, introduce the *extended twisted Yangian* $\mathrm{X}(\mathfrak{gl}_N, \sigma)$. The unital associative algebra $\mathrm{X}(\mathfrak{gl}_N, \sigma)$ has a family of generators $S_{ij}^{(a)}$ where $a = 1, 2, \ldots$ and $i, j = 1, \ldots, N$. Put

$$S_{ij}(x) = \delta_{ij} \cdot 1 + S_{ij}^{(1)} x^{-1} + S_{ij}^{(2)} x^{-2} + \ldots \in \mathrm{X}(\mathfrak{gl}_N, \sigma)\,[[x^{-1}]]\,. \tag{2.2}$$

Defining relations for the generators $S_{ij}^{(a)}$ of the algebra $\mathrm{X}(\mathfrak{gl}_N, \sigma)$ can be written as

$$(x^2 - y^2) \cdot [S_{ij}(x), S_{kl}(y)] = (x + y) \cdot \big(S_{kj}(x) S_{il}(y) - S_{kj}(y) S_{il}(x)\big)$$

$$- (x - y) \cdot \big(S_{ik}(x) S_{jl}(y) - S_{ki}(y) S_{lj}(x)\big) + S_{ki}(x) S_{jl}(y) - S_{ki}(y) S_{jl}(x)\,.$$

All these relations can be written as a single *reflection equation*, see [7]. One can define a homomorphism $\pi_N : \mathrm{X}(\mathfrak{gl}_N, \sigma) \to \mathrm{Y}(\mathfrak{gl}_N, \sigma)$ by mapping the series $S_{ij}(x)$ to (2.1). The homomorpism π_N is surjective. As a two-sided ideal of $\mathrm{X}(\mathfrak{gl}_N, \sigma)$, the kernel of the homomorphism π_N is generated by the coefficients of all series

$$S_{ij}(x) + (2x - 1)\, S_{ij}(-x) - 2x\, S_{ji}(x) \tag{2.3}$$

in x^{-1}. This ideal is also generated by certain central elements of $\mathrm{X}(\mathfrak{gl}_N, \sigma)$, see [7].

The algebra $\mathrm{X}(\mathfrak{gl}_N, \sigma)$ admits an analogue of the automorphism ξ_N of $\mathrm{Y}(\mathfrak{gl}_N)$. Determine a formal power series $\widetilde{S}_{ij}(x)$ in x^{-1} with the coefficients in $\mathrm{X}(\mathfrak{gl}_N, \sigma)$ and the leading term δ_{ij}, by the system of equations

$$\sum_{k=1}^{N} S_{ik}(x)\, \widetilde{S}_{kj}(x) = \delta_{ij} \quad \text{where} \quad i, j = 1, \ldots, N.$$

Then one can define an involutive automorphism η_N of the algebra $\mathrm{X}(\mathfrak{gl}_N,\sigma)$ by the assignment

$$\eta_N : S_{ij}(x) \mapsto \widetilde{S}_{ij}(-x-\tfrac{N}{2}).$$

However, η_N does not determine an automorphism of $\mathrm{Y}(\mathfrak{gl}_N,\sigma)$, because η_N does not preserve the ideal of $\mathrm{X}(\mathfrak{gl}_N,\sigma)$ generated by the coefficients of all series (2.3).

For any formal power series $f(x) \in \mathbb{C}[[x^{-1}]]$ with the leading term 1, the assignment

$$S_{ij}(x) \mapsto f(x)\cdot S_{ij}(x) \tag{2.4}$$

defines an automorphism of the algebra $\mathrm{X}(\mathfrak{gl}_N,\sigma)$. The defining relations of the algebra $\mathrm{X}(\mathfrak{gl}_N,\sigma)$ imply that the assignment

$$\beta_N : S_{ij}(x) \mapsto \delta_{ij}\cdot 1 + \frac{E_{ij}-E_{ji}}{x+\frac{1}{2}}$$

defines a homomorphism of associative algebras $\beta_N : \mathrm{X}(\mathfrak{gl}_N,\sigma) \to \mathrm{U}(\mathfrak{so}_N)$. By definition, the homomorphism β_N is surjective. Moreover, β_N factors through π_N. Note that the homomorphism $\mathrm{Y}(\mathfrak{gl}_N,\sigma) \to \mathrm{U}(\mathfrak{so}_N)$ corresponding to β_N, cannot be obtained from $\alpha_N : \mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{U}(\mathfrak{gl}_N)$ by restricting to the subalgebra $\mathrm{Y}(\mathfrak{gl}_N,\sigma)$, because the image of $\mathrm{Y}(\mathfrak{gl}_N,\sigma)$ relative to α_N is not contained in the subalgebra $\mathrm{U}(\mathfrak{so}_N) \subset \mathrm{U}(\mathfrak{gl}_N)$; see [11]. An embedding $\mathrm{U}(\mathfrak{so}_N) \to \mathrm{Y}(\mathfrak{gl}_N,\sigma)$ can be defined by

$$E_{ij}-E_{ji} \mapsto T^{(1)}_{ij} - T^{(1)}_{ji},$$

cf. (1.12). The homomorphism $\mathrm{Y}(\mathfrak{gl}_N,\sigma) \to \mathrm{U}(\mathfrak{so}_N)$ corresponding to β_N, is then identical on the subalgebra $\mathrm{U}(\mathfrak{so}_N) \subset \mathrm{Y}(\mathfrak{gl}_N,\sigma)$.

2.2. For any partition ν with $\nu'_1 \leqslant N$, the irreducible polynomial $\mathfrak{gl}_N$-module V_ν can also be regarded as a representation of the complex general linear Lie group GL_N. Consider the subgroup $O_N \subset GL_N$ preserving the standard symmetric bilinear form $\langle\ ,\ \rangle$ on $\mathbb{C}^N$. The subalgebra $\mathfrak{so}_N \subset \mathfrak{gl}_N$ corresponds to this subgroup. Note that the complex Lie group O_N has two connected components. In [20] the irreducible finite-dimensional representations of the group O_N are labeled by the partitions ν of $n = 0,1,2,\ldots$ such that $\nu'_1+\nu'_2 \leqslant N$. Denote by W_ν the irreducible representation of O_N corresponding to ν. As $\mathfrak{so}_N$-module, W_ν is irreducible unless $2\nu'_1 = N$, in which case W_ν is a direct sum of two irreducible $\mathfrak{so}_N$-modules.

Choose any embedding of the irreducible representation V_ν of the group GL_N into the space $(\mathbb{C}^N)^{\otimes n}$. Take any two distinct numbers $k,l \in \{1,\ldots,n\}$. By applying the bilinear form $\langle\ ,\ \rangle$ to a tensor $w \in (\mathbb{C}^N)^{\otimes n}$ in the kth and lth tensor factors, we obtain a certain tensor $\widehat{w} \in (\mathbb{C}^N)^{\otimes(n-2)}$. The tensor w is called *traceless*, if $\widehat{w} = 0$ for all distinct k and l. Denote by $(\mathbb{C}^N)^{\otimes n}_0$ the subspace in $(\mathbb{C}^N)^{\otimes n}$ consisting of all traceless tensors, this subspace is O_N-invariant. Then W_ν can be embedded into $(\mathbb{C}^N)^{\otimes n}$ as the intersection $V_\nu \cap (\mathbb{C}^N)^{\otimes n}_0$, see [20].

Let the indices i and j range over $\{1,\ldots,N+M\}$. Choose the embedding of the Lie algebras $\mathfrak{gl}_N$ and $\mathfrak{gl}_M$ into $\mathfrak{gl}_{N+M}$ as in Subsection 1.2. It determines embeddings of groups $GL_N \times GL_M \to GL_{N+M}$ and $O_N \times O_M \to O_{N+M}$. Take any two partitions λ and μ such that $\lambda'_1+\lambda'_2 \leqslant N+M$ and $\mu'_1+\mu'_2 \leqslant M$. Consider the

irreducible representations W_λ and W_μ of the groups O_{N+M} and O_M respectively. The vector space

$$\mathrm{Hom}_{O_M}(W_\mu, W_\lambda) \tag{2.5}$$

comes with a natural action of the group O_N. This action of O_N may be reducible. The vector space (2.5) is non-zero, if and only if $\lambda_k \geqslant \mu_k$ and $\lambda'_k - \mu'_k \leqslant N$ for each $k = 1, 2, \ldots$; see [18]. Thus for a given N, the vector spaces (1.5) and (2.5) are zero or non-zero simultaneously. Further, for a given N, the dimension of (2.5) does not exceed that of (1.5). Our results provide an embedding of (2.5) into (1.5), compatible with the action of the orthogonal group O_N in these two vector spaces.

Denote by $\mathrm{B}_N(M)$ the subalgebra of O_M-invariants in the universal enveloping algebra $\mathrm{U}(\mathfrak{so}_{N+M})$. Then $\mathrm{B}_N(M)$ contains the subalgebra $\mathrm{U}(\mathfrak{so}_N) \subset \mathrm{U}(\mathfrak{so}_{N+M})$, and is contained in the centralizer of the subalgebra $\mathrm{U}(\mathfrak{so}_M) \subset \mathrm{U}(\mathfrak{so}_{N+M})$. The algebra $\mathrm{B}_N(M)$ naturally acts in the vector space (2.5). The $\mathrm{B}_N(M)$-module (2.5) is either irreducible, or splits into a direct sum of two irreducible $\mathrm{B}_N(M)$-modules. In the latter case, (2.5) is irreducible under the joint action of the algebra $\mathrm{B}_N(M)$ and the subgroup $O_N \subset O_{N+M}$.

For every non-negative integer M, Olshanski [17] defined a homomorphism $\mathrm{Y}(\mathfrak{gl}_N, \sigma) \to \mathrm{B}_N(M)$. Along with the subalgebra of O_{N+M}-invariants in $\mathrm{U}(\mathfrak{so}_{N+M})$, the image of this homomorphism generates the algebra $\mathrm{B}_N(M)$. We use a version of this homomorphism for the algebra $\mathrm{X}(\mathfrak{gl}_N, \sigma)$, this version is denoted by β_{NM}.

Consider the extended twisted Yangian $\mathrm{X}(\mathfrak{gl}_{N+M}, \sigma)$, where $-\sigma$ is the matrix transposition in $\mathfrak{gl}_{N+M}$. The subalgebra in $\mathrm{X}(\mathfrak{gl}_{N+M}, \sigma)$ generated by $S_{ij}^{(a)}$ where $i, j = 1, \ldots, N$, by definition coincides with $\mathrm{X}(\mathfrak{gl}_N, \sigma)$. Denote by ψ_M this natural embedding $\mathrm{X}(\mathfrak{gl}_N, \sigma) \to \mathrm{X}(\mathfrak{gl}_{N+M}, \sigma)$. Consider also the involutive automorphism η_{N+M} of the algebra $\mathrm{X}(\mathfrak{gl}_{N+M}, \sigma)$. The image of the homomorphism

$$\beta_{N+M} \circ \eta_{N+M} \circ \psi_M : \mathrm{X}(\mathfrak{gl}_N, \sigma) \to \mathrm{U}(\mathfrak{so}_{N+M})$$

belongs to the subalgebra $\mathrm{B}_N(M) \subset \mathrm{U}(\mathfrak{so}_{N+M})$. Moreover, this image along with the subalgebra of O_{N+M}-invariants in $\mathrm{U}(\mathfrak{so}_{N+M})$, generates $\mathrm{B}_N(M)$; see [9]. We use the homomorphism $\mathrm{X}(\mathfrak{gl}_N, \sigma) \to \mathrm{U}(\mathfrak{so}_{N+M})$

$$\beta_{NM} = \beta_{N+M} \circ \eta_{N+M} \circ \psi_M \circ \eta_N . \tag{2.6}$$

When $M = 0$, the homomorphism (2.6) coincides with β_N. The intersection of the kernels of all homomorphisms $\beta_{N0}, \beta_{N1}, \beta_{N2}, \ldots$ is contained in the kernel of π_N.

2.3. The $\mathrm{B}_N(M)$-module (2.5) depends on the partitions λ and μ via the skew Young diagram ω. Using the homomorphism $\beta_{NM} : \mathrm{X}(\mathfrak{gl}_N, \sigma) \to \mathrm{B}_N(M)$, regard (2.5) as $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-module. Unlike the $\mathrm{Y}(\mathfrak{gl}_N)$-module $V_\omega(z)$, this $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-module may depend on the choice of the integer M, such that $\lambda'_1 + \lambda'_2 \leqslant N + M$ and $\mu'_1 + \mu'_2 \leqslant M$. Denote this $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-module by $W_\omega(M)$. Note that when $z \neq 0$, the automorphism τ_z of $\mathrm{Y}(\mathfrak{gl}_N)$ does not preserve the subalgebra $\mathrm{Y}(\mathfrak{gl}_N, \sigma) \subset \mathrm{Y}(\mathfrak{gl}_N)$. There is no analogue of the automorphism τ_z with $z \neq 0$ for the algebra $\mathrm{X}(\mathfrak{gl}_N, \sigma)$.

The O_{N+M}-invariant elements of $\mathrm{U}(\mathfrak{so}_{N+M})$ act in (2.5) as scalar operators. Thus the $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-module $W_\omega(M)$ is either irreducible, or splits into a direct sum of two irreducible $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-modules. In the latter case, it becomes irreducible

under the joint action of the algebra $\mathrm{X}(\mathfrak{gl}_N,\sigma)$ and the subgroup $O_N \subset O_{N+M}$. Our main result is an explicit realization of the $\mathrm{X}(\mathfrak{gl}_N,\sigma)$-module $W_\omega(M)$, similar to the realization of the elementary $\mathrm{Y}(\mathfrak{gl}_N)$-module given by Theorem 2. Our explicit realization is compatible with the action of the group O_N in $W_\omega(M)$.

Take the standard orthonormal basis $e_1, \ldots, e_N$ in $\mathbb{C}^N$, so that $\langle e_i, e_j \rangle = \delta_{ij}$. The linear operator

$$u \otimes v \mapsto \langle u,v \rangle \cdot \sum_{i=1}^{N} e_i \otimes e_i \tag{2.7}$$

in $\mathbb{C}^N \otimes \mathbb{C}^N$ commutes with the action of O_N. Take the complex variables $t_1, \ldots, t_n$ with the same constraints as in Proposition 1. Consider the ordered product over the pairs (k,l),

$$\prod_{1 \leqslant k < l \leqslant n}^{\longrightarrow} \left(1 - \frac{Q_{kl}}{c_k + c_l + t_k + t_l + N + M}\right) \tag{2.8}$$

where Q_{kl} is the linear operator in $(\mathbb{C}^N)^{\otimes n}$, acting as (2.7) in the kth and lth tensor factors, and acting as the identity in the remaining $n-2$ tensor factors. Here the pairs (k,l) are ordered lexicographically, as in (1.9). Let us now multiply (2.8) by (1.9) on the right, and consider the result as an operator-valued rational function of the constrained variables $t_1, \ldots, t_n$.

Proposition 3. *At $t_1 = \ldots = t_n = -\frac{1}{2}$ the ordered product of* (2.8) *and* (1.9) *has the value*

$$\prod_{(k,l)}^{\longrightarrow} \left(1 - \frac{Q_{kl}}{c_k + c_l + N + M - 1}\right) \cdot E_\Omega = E_\Omega \cdot \prod_{(k,l)}^{\longleftarrow} \left(1 - \frac{Q_{kl}}{c_k + c_l + N + M - 1}\right); \tag{2.9}$$

the ordered products in (2.9) *are taken over all pairs (k,l) such that the numbers k and l appear in different columns of the tableau Ω.*

Denote the operator (2.9) by $F_\Omega(M)$. If k and l appear in different columns of Ω, then

$$c_k + c_l \geqslant 3 - \lambda'_1 - \lambda'_2 \geqslant 3 - N - M\,.$$

Hence each of the denominators in (2.9) is non-zero for any choice of μ. The algebra of operators in $(\mathbb{C}^N)^{\otimes n}$ generated by all P_{kl} and Q_{kl} with $1 \leqslant k < l \leqslant n$, is called the *Brauer centralizer algebra*; see [1]. The operator $F_\Omega(M)$ belongs to this algebra. Note that the image of the operator $F_\Omega(M)$ is contained in the image of E_Ω.

Suppose that $M = 0$, then $\mu = (0,0,\ldots)$. In this special case, the image of the operator E_Ω in $(\mathbb{C}^N)^{\otimes n}$ is equivalent to V_λ as a representation of the group GL_N, see (1.10). It turns out that the image of the operator $F_\Omega(0)$ consists of all traceless tensors from the image of E_Ω. In particular, the image of $F_\Omega(0)$ in $(\mathbb{C}^N)^{\otimes n}$ is equivalent to W_λ as a representation of the group O_N. Even in the special case $M = 0$, the formulas (2.9) for the operator $F_\Omega(M)$ seem to be new; cf. [20].

2.4. Let us extend σ to an automorphism of the associative algebra $\mathrm{U}(\mathfrak{gl}_N)$. For any $z \in \mathbb{C}$, define the *twisted evaluation module* $\widetilde{V}(z)$ over the algebra $\mathrm{Y}(\mathfrak{gl}_N)$ by pulling the standard action of the algebra $\mathrm{U}(\mathfrak{gl}_N)$ in the vector space $\mathbb{C}^N$ back through the composition of homomorphisms

$$\sigma \circ \alpha_N \circ \tau_{-z} : \mathrm{Y}(\mathfrak{gl}_N) \to \mathrm{U}(\mathfrak{gl}_N)\,.$$

The evaluation module $V(z)$ and the twisted evaluation module $\widetilde{V}(z)$ over $\mathrm{Y}(\mathfrak{gl}_N)$, have the same restriction to the subalgebra $\mathrm{Y}(\mathfrak{gl}_N, \sigma) \subset \mathrm{Y}(\mathfrak{gl}_N)$; see (2.1).

For any λ and μ, the operator $F_\Omega(M)$ has the following interpretation in terms of the restrictions to $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ of tensor products of evaluation modules over the Hopf algebra $\mathrm{Y}(\mathfrak{gl}_N)$; cf. Proposition 2. For each $k = 1, \ldots, n$ put $d_k = c_k + \frac{M}{2} - \frac{1}{2}$. We assume that $\lambda'_1 + \lambda'_2 \leqslant N + M$ and $\mu'_1 + \mu'_2 \leqslant M$.

Proposition 4. *The operator $F_\Omega(M)$ is an intertwiner of $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$-modules*

$$\widetilde{V}(d_1) \otimes \ldots \otimes \widetilde{V}(d_n) \longrightarrow V(d_1) \otimes \ldots \otimes V(d_n)\,.$$

By Proposition 4, the image of $F_\Omega(M)$ is a submodule in the restriction of the tensor product of evaluation $\mathrm{Y}(\mathfrak{gl}_N)$-modules $V(d_1) \otimes \ldots \otimes V(d_n)$ to the subalgebra $\mathrm{Y}(\mathfrak{gl}_N, \sigma) \subset \mathrm{Y}(\mathfrak{gl}_N)$. Denote this $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$-submodule by $W_\Omega(M)$. It is also a submodule in the restriction of the $\mathrm{Y}(\mathfrak{gl}_N)$-module $V_\Omega(\frac{M}{2} - \frac{1}{2})$ to $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$.

Theorem 3. *a) By pulling the $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-module $W_\omega(M)$ back through the automorphism of $\mathrm{X}(\mathfrak{gl}_N, \sigma)$ defined by (2.4) where $f(x) = f_\mu(x - \frac{M}{2} + \frac{1}{2})$, we get an $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-module that factors through homomorphism $\pi : \mathrm{X}(\mathfrak{gl}_N, \sigma) \to \mathrm{Y}(\mathfrak{gl}_N, \sigma)$.*

b) This $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$-module, corresponding to $W_\omega(M)$, is equivalent to $W_\Omega(M)$.

The vector space (2.5) of the $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-module $W_\omega(M)$ comes with a natural action of the group O_N. The action of the group O_N in $(\mathbb{C}^N)^{\otimes n}$ preserves the image of the operator $F_\Omega(M)$, because $F_\Omega(M)$ commutes with this action. Thus the vector space of the $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$-module $W_\Omega(M)$ also comes with an action of O_N. The proof of Theorem 3 is given in [12]. It provides an O_N-equivariant isomorphism between the $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$-module corresponding to $W_\omega(M)$, and the $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$-module $W_\Omega(M)$. This isomorphism is unique, up to a scalar multiplier. The image of the operator $F_\Omega(M)$ is irreducible under the joint action of $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$ and O_N.

Thus we can identify the vector space (2.5) with the image of the operator $F_\Omega(M)$ uniquely, up to multiplication in (2.5) by a non-zero complex number. Using Theorem 2, we can identify the vector space (1.5) with the image of E_Ω, again uniquely up to rescaling. Since the image of $F_\Omega(M)$ is contained in that of E_Ω, we then obtain a distinguished embedding of the vector space (2.5) into (1.5).

Theorem 3 provides an explicit realization of the $\mathrm{X}(\mathfrak{gl}_N, \sigma)$-module $W_\omega(M)$ as a subspace in $(\mathbb{C}^N)^{\otimes n}$. This theorem also turns the vector space (2.5) into a module over the twisted Yangian $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$, equivalent to $W_\Omega(M)$. The limited size of the present article does not allow us to discuss here the analogues of the results of [14] and [15] for the $\mathrm{Y}(\mathfrak{gl}_N, \sigma)$-modules, obtained in this particular way; cf. [13].

References

[1] R. Brauer, On algebras which are connected with the semisimple continuous groups, *Ann. Math.*, 38 (1937), 857–872.

[2] I. Cherednik, Factorized particles on the half-line and root systems, *Theor. Math. Phys.*, 61 (1984), 977–983.

[3] I. Cherednik, A new interpretation of Gelfand-Zetlin bases, *Duke Math. J.*, 54 (1987), 563–577.

[4] V. Drinfeld, Hopf algebras and the quantum Yang–Baxter equation, *Soviet Math. Doklady*, 32 (1985), 254–258.

[5] V. Drinfeld, A new realization of Yangians and quantized affine algebras, *Soviet Math. Doklady*, 36 (1988), 212–216.

[6] P. Kulish and E. Sklyanin, Quantum spectral transform method: recent developments, *Lecture Notes in Phys.*, 151 (1982), 61–119.

[7] A. Molev, M. Nazarov and G. Olshanski, Yangians and classical Lie algebras, *Russian Math. Surveys*, **51** (1996), 205–282.

[8] I. Macdonald, *Symmetric Functions and Hall Polynomials*, Clarendon Press, 1995.

[9] A. Molev and G. Olshanski, Centralizer construction for twisted Yangians, *Selecta Math.*, 6 (2000), 269–317.

[10] M. Nazarov, Yangians and Capelli identities, *Amer. Math. Soc. Translations*, 181 (1998), 139–163.

[11] M. Nazarov, Capelli elements in the classical universal enveloping algebras, *Adv. Stud. Pure Math.*, 28 (2000), 261–285.

[12] M. Nazarov, Representations of twisted Yangians associated with skew Young diagrams, *in preparation*.

[13] M. Nazarov and G. Olshanski, Bethe subalgebras in twisted Yangians, *Comm. Math. Phys.*, 178 (1996), 483–506.

[14] M. Nazarov and V. Tarasov, Representations of Yangians with Gelfand-Zetlin bases, *J. Reine Angew. Math.*, 496 (1998), 181–212.

[15] M. Nazarov and V. Tarasov, On irreducibility of tensor products of Yangian modules associated with skew Young diagrams, *Duke Math. J.*, 112 (2002), 342–378.

[16] G. Olshanski, Extension of the algebra $U(g)$ for infinite-dimensional classical Lie algebras g, and the Yangians $Y(gl(m))$, *Soviet Math. Doklady*, 36 (1988), 569–573.

[17] G. Olshanski, Twisted Yangians and infinite-dimensional classical Lie algebras, *Lecture Notes in Math.*, 1510 (1992), 103–120.

[18] R. Proctor, *Young tableaux, Gelfand patterns, and branching rules for classical groups*, J. Algebra **164** (1994), 299–360.

[19] E. Sklyanin, Boundary conditions for integrable quantum systems, *J. Phys.*, Series A, 21 (1988), 2375–2389.

[20] H. Weyl, *Classical Groups, their Invariants and Representations*, Princeton University Press, 1946.

ICM 2002 · Vol. II · 655–666

Automorphic L-Functions and Functoriality

Freydoon Shahidi*

Abstract

This is a report on the global aspects of the Langlands-Shahidi method which in conjunction with converse theorems of Cogdell and Piatetski-Shapiro has recently been instrumental in establishing a significant number of new and surprising cases of Langlands Functoriality Conjecture over number fields. They have led to striking new estimates towards Ramanujan and Selberg conjectures.

2000 Mathematics Subject Classification: 11F70, 11R39, 11R42, 11S37, 22E55.
Keywords and Phrases: Automorphic L-function, Functoriality.

1. Preliminaries

Let F be a number field. For each place v of F, let F_v be its completion at v. Assume v is a finite place and let O_v denote the ring of integers of F_v. Denote by P_v its maximal ideal and fix a uniformizing parameter ϖ_v generating P_v. Let $[O_v : P_v] = q_v$ and fix and absolute value $|\quad|_v$ for which $|\varpi_v|_v = q_v^{-1}$.

Let $\mathbf{G}$ be a quasisplit connected reductive algebraic group over F. Fix an F-Borel subgroup $\mathbf{B} = \mathbf{TU}$, where $\mathbf{T}$ is a maximal torus of $\mathbf{B}$ and $\mathbf{U}$ is its unipotent radical. Let $\mathbf{A_0} \subset \mathbf{T}$ be the maximal split subtorus of $\mathbf{T}$. Throughout this article, $\mathbf{P}$ is a maximal parabolic subgroup of $\mathbf{G}$, defined over F, with a Levi decomposition $\mathbf{P} = \mathbf{MN}$, where $\mathbf{M}$ is a Levi subgroup of $\mathbf{P}$ and $\mathbf{N}$ is its unipotent radical. We will assume $\mathbf{P}$ is standard in the sense that $\mathbf{N} \subset \mathbf{U}$. We fix $\mathbf{M}$ by assuming $\mathbf{T} \subset \mathbf{M}$. We finally use W to denote the Weyl group of $\mathbf{A_0}$ in $\mathbf{G}$.

Let $\mathbb{A}_F$ denote the ring of adeles of F and for every algebraic group $\mathbf{H}$ over F, let $H = \mathbf{H}(\mathbb{A}_F)$. Considering $\mathbf{H}$ as a group over each F_v, we then set $H_v = \mathbf{H}(F_v)$.

Let $\mathbf{A}$ denote the split component of $\mathbf{M}$, i.e., the maximal split subtorus of the connected component of the center of $\mathbf{M}$. For every group $\mathbf{H}$ defined over F, let

*Purdue University, Department of Mathematics, West Lefayette, Indiana 47907, USA. E-mail: shahidi@math.purdue.edu

$X(\mathbf{H})_F$ be the group of F-rational characters of $\mathbf{H}$. We set $\mathfrak{a} = \mathrm{Hom}(X(\mathbf{M})_F, \mathbb{R})$. Then $\mathfrak{a}^* = X(\mathbf{M})_{F\otimes\mathbb{Z}}\mathbb{R} = X(\mathbf{A})_{F\otimes\mathbb{Z}}\mathbb{R}$ and $\mathfrak{a}_{\mathbb{C}}^* = \mathfrak{a}^* \otimes_{\mathbb{R}} \mathbb{C}$ is the complex dual of $\mathfrak{a}$.

When $\mathbf{G}$ is unramified over a place v, we let $K_v = \mathbf{G}(\mathbf{O}_v)$. Otherwise, we shall fix a special maximal compact subgroup $K_v \subset G_u$ for which $G_v = P_v K_v = B_v K_v$. Let $K = \otimes_v K_v$ Then $G = PK = BK$. Let $K_M = K \cap M$.

For each v, the embedding $X(\mathbf{M})_F \hookrightarrow X(\mathbf{M})_{F_v}$ induces a map

$$a_v = \mathrm{Hom}(X(\mathbf{M})_{F_v}, \mathbb{R}) \to \mathfrak{a}.$$

There exists a homomorphism $H_M : M \to \mathfrak{a}$ defined by

$$\exp\langle \chi, H_M(m)\rangle = \prod_v |\chi(m_v)|_v$$

for every $\chi \in X(\mathbf{M})_F$ and $m = (m_v)$. We extend H_M to H_P on G by making it trivial on N and K.

Let α denote the unique simple root of $\mathbf{A}$ in $\mathbf{N}$. It can be identified by a unique simple root of $\mathbf{A_0}$ in $\mathbf{U}$. If $\rho_{\mathbf{P}}$ is half the sum of F-roots in $\mathbf{N}$, we set $\widetilde{\alpha} = \langle \rho_{\mathbf{P}}, \alpha\rangle^{-1}\rho_{\mathbf{P}} \in a^*$, where for each pair of non-restricted roots α and β of $\mathbf{T}$, $\langle \alpha, \beta\rangle = 2(\alpha,\beta)/(\beta,\beta)$ is the Killing form.

Given a connected reductive algebraic group $\mathbf{H}$ over F, let ${}^L H$ be its L-group. Considering $\mathbf{H}$ as a group over F_v, we then denote by ${}^L H_v$ its L-group over F_v. Let ${}^L H^0 = {}^L H_v^0$ be the corresponding connected component of 1. We then have a natural homomorphism from ${}^L H_v$ into ${}^L H$. We let $\eta_v : {}^L M_v \to {}^L M$ be this map for M (cf. [4]).

Let ${}^L N$ be the L-group of $\mathbf{N}$ defined naturally in [4]. Let ${}^L\mathfrak{n}$ be its (complex) Lie algebra, and let r denote the adjoint action of ${}^L M$ on ${}^L\mathfrak{n}$. Decompose $r = \bigoplus_{i=1}^{m} r_i$ to its irreducible subrepresentations, indexed according to the values $\langle \widetilde{\alpha}, \beta\rangle = i$ as β ranges among the positive roots of $\mathbf{T}$. More precisely, $X_{\beta\vee} \in {}^L\mathfrak{n}$ lies in the space of r_i if and only if $\langle \widetilde{\alpha}, \beta\rangle = i$. Here $X_{\beta\vee}$ is a root vector attached to the coroot $\beta^\vee$, considered as a root of the L-group. The integer m is equal to the nilpotence class of ${}^L\mathfrak{n}$. We let $r_{i,v} = r_i \cdot \eta_v$ for each i (cf. [34,40,41]).

If Δ denotes the set of simple roots of $\mathbf{A}_0$ in $\mathbf{U}$, we use $\theta \subset \Delta$ to denote the subset generating M. Then $\Delta = \theta \cup \{\alpha\}$. There exists a unique element $\widetilde{w}_0 \in W$ such that $\widetilde{w}_0(\theta) \subset \Delta$, while $\widetilde{w}_0(\alpha) < 0$. We will always choose a representative w_0 for $\widetilde{w}_0$ in $\mathbf{G}(F)$ and use w_0 to denote each of its components.

2. Eisenstein series and L-functions

Let $\pi = \otimes_v \pi_v$ be a cusp form on M. Given a K_M-finite function φ in the space of π, we extend φ to a function $\widetilde{\varphi}$ on G as in Section 2 of [39] as well as in [17], and for $s \in \mathbb{C}$, set

$$\phi_s(g) = \widetilde{\varphi}(g)\exp\langle s\widetilde{\alpha} + \rho_{\mathbf{P}}, H_P(g)\rangle. \tag{2.1}$$

The corresponding Eisenstein series is then defined by

$$E(s,\phi_s,g,P)=\sum_{\gamma\in\mathbf{P}(F)\backslash\mathbf{G}(F)}\phi_s(\gamma g) \tag{2.2}$$

(cf. [17,33,34,35]).

Let $I(s,\pi)=\otimes_v I(s,\pi_v)$ be the representation parabolically induced from $\pi\otimes\exp\langle s\widetilde{\alpha},H_p(\,)\rangle$.

Let $\mathbf{M}'$ be the Levi subgroup of $\mathbf{G}$ generated by $\widetilde{w}(\theta)$. There exists a parabolic subgroup $\mathbf{P}'\supset\mathbf{B}$ which has $\mathbf{M}'$ as a Levi factor. Let $\mathbf{N}'$ be its unipotent radical. Given f in the space of $I(s,\pi)$ and $\mathrm{Re}(s)>>0$, define the global intertwining operator $M(s,\pi)$ by

$$M(s,\pi)f(g)=\int_{N'}f(w_0^{-1}n'g)dn' \quad (g\in G). \tag{2.3}$$

Observe that if $f=\otimes_v f_v$, then for almost all v, f_v is the unique K_v-fixed functions normalized by $f_v(e_v)=1$. Finally, if at each v we define a local intertwining operator by

$$A(s,\pi_v,w_0)f_v(g)=\int_{N'_v}f_v(w_0^{-1}n'g)dn', \tag{2.4}$$

then

$$M(s,\pi)=\otimes_v A(s,\pi_v,w_0). \tag{2.5}$$

It follows form the general theory of Eisenstein series that the poles of $E(s,\widetilde{\varphi},g,P)$, as $\widetilde{\varphi}$ and g vary, are the same as those of $M(s,\pi)$, and for $\mathrm{Re}(s)\geq 0$, they are all simple and finite in number, with none on the line $\mathrm{Re}(s)=0$ (cf. [17,33,35]).

By construction each ϕ_s belongs to the space of $I(s,\pi)$. Consequently, one can consider $M(s,\pi)\phi_s$ which is a member of $I(-s,w_0(\pi))$. The Eisenstein series $E(s,\widetilde{\varphi},g,P)$ then satisfies the functional equation

$$E(s,\phi_s,g,P)=E(-s,M(s,\pi)\phi_s,g,P'). \tag{2.6}$$

Suppose that $\mathbf{G}$ splits over L, where L is a finite Galois extension of F. For every unramified v, there exists a unique Frobenius conjugacy class in $\mathrm{Gal}(L_w/F_v)$, $w|v$ which we denote by τ_v. Moreover, if v is such that π_v and $\mathbf{G}$ are both unramified, then there exists and ${}^LM$ semisimple conjugacy class in ${}^LM^0\rtimes\tau_v$ which determines π_v uniquely ([4,40]). We may identify, as we in fact do, this conjugacy class with an element $A_v\in{}^LT^0$ which may be assumed to be fixed by τ_v (cf. §6.3 and 6.5 of [4]). The local Langlands L-function defined by π_v and $r_v, r_v=r\cdot\eta_v$, where r is a complex analytic representation of ${}^LM$, is then defined to be (cf. [4,34,40]),

$$L(s,\pi_v,r_v)=\det(I-r_v(A_v\rtimes\tau_v)q_v^{-1})^{-1}. \tag{2.7}$$

Let S be a finite set of places of F, including all the archimedean ones, such that for every $v\notin S$, π_v and $\mathbf{G}$ are both unramified. Set

$$L_S(s,\pi,r)=\prod_{v\notin S}L(s,\pi_v,r_v). \tag{2.8}$$

The main result of [34, also see 40] is that

$$M(s,\pi)f = \otimes_{v\in S}A(s,\pi_v,w_0)f_v \otimes \otimes_{v\notin S}\widetilde{f}_v \times \prod_{i=1}^{m} L_S(is,\pi,\widetilde{r}_i)/L_S(1+is,\pi,\widetilde{r}_i), \tag{2.9}$$

where $f = \otimes_v f_v$ is such that for each $v \notin S$, f_v is the unique K_v-fixed function in $I(s,\pi_v)$ normalized by $f_v(e_v) = 1$ and for each i, $\widetilde{r}_i$ denotes the contragredient of $r_i, i = 1,\cdots,m$, the irreducible components of the adjoint action of ${}^LM$ or ${}^LN$. Here $\widetilde{f}_v$ is the K_v-fixed function in the space of $I(-s,w_0(\pi_v))$, normalized the same way. Moreover f_v and $\widetilde{f}_v$ are identified as elements in spherical principal series.

3. Generic representations and the non-constant term

Suppose F is a field, either local or global, and $\mathbf{G}$ is as before, with a Borel subgroup $\mathbf{B} = \mathbf{TU}$ over F. Fix an F-splitting $\{X_{\alpha'}\}$, i.e., a collection of root vectors as α' ranges over simple roots of $\mathbf{T}$ in $\mathbf{U}$ which is invariant under the action of $\Gamma_F = \mathrm{Gal}(\overline{F}/F)$. This then determines a map ϕ form $\mathbf{U}$ to $\Pi\mathbb{G}_a, \varphi(u) = (x_{\alpha'})_{\alpha'}$, where $x_{\alpha'}$ is the α'-coordinate of u with respect to $\{X_{\alpha'}\}$. Let $\{\kappa_{\alpha'}\}$ be a collection of elements in $\overline{F}^*$ such that $\sigma(\kappa_{\alpha'}) = \kappa_{\sigma\alpha'}$ for every $\sigma \in \Gamma_F$. Set $f(u) = \sum_{\alpha'} \kappa_{\alpha'} x_{\alpha'}$. Observe that f is F-rational. If F is global, we extend f to a map on $\mathbf{U}(\mathbb{A}_F)$. Let ψ_F be a non-trivial character of F ($F \setminus \mathbb{A}_F$ if F is global). A character χ of $\mathbf{U}(F)(\mathbf{U}(F) \setminus \mathbf{U}(\mathbb{A}_F)$ if F is global) is called *non − degenerate* or *generic* if $\chi(u) = \varphi(f(u)), u \in \mathbf{U}(F)(u \in \mathbf{U}(F) \setminus \mathbf{U}(\mathbb{A}_F)$ if F is global).

We now continue to assume F is a number field. Let $\chi = \otimes_v \chi_v$ be a generic character of $\mathbf{U}(F) \setminus U$.

Let $\mathbf{U}^0 = \mathbf{U} \cap \mathbf{M}$ and let χ also denote the restriction of χ to U^0. Choose a function φ in the space of $\pi = \otimes_v \pi_v$, a cuspidal representation of M, and $\mathbf{U}^0(F)\backslash U^0$ being compact, set

$$W_\varphi(m) = \int_{\mathbf{U}^0(F)\backslash U^0} \varphi(um)\overline{\chi(u)}du. \tag{3.1}$$

We shall say π is (globally) χ-generic if $W_\varphi \neq 0$ for some φ. The representation π is (globally) generic if it is χ-generic with respect to some generic χ. Then each π_v will be χ_v-generic in the sense that there exists a non-zero Whittaker functional λ_v i.e., a continuous (in the semi-norm topology if $v = \infty$) functional satisfying $\langle \pi_v(u)x, \lambda_v\rangle = \chi_v(u)\langle x, \lambda_v\rangle, x \in \mathcal{H}(\pi_v), u \in U_v^0$. Choosing φ appropriately, i.e., if $\varphi = \otimes_v \varphi_v, \varphi_v \in \mathcal{H}(\pi_v)$, then $W_\varphi(m) = \prod_v \langle \pi_v(m_v)\varphi_v, \lambda_v\rangle$, for $m = (m_v)$.

Given $f_v \in V(s,\pi_v)$, the space of $I(s,\pi_v)$, define

$$\lambda_{\lambda_v}(s,\pi_v)(f_v) = \int_{N_v'} \langle f_v(w_0^{-1}n'), \lambda_v\rangle\overline{\chi(n')}dn', \tag{3.2}$$

a canonical Whittaker functional for $I(s,\pi_v)$. Changing the splitting we now assume $\kappa_{\alpha'}=1$. It now follows from Rodier's theorem that there exists a complex function (of s), $C_{\chi_v}(s,\pi_v)$, depending on π_v,χ_v and w_0 such that (cf. [41,42,43])

$$\lambda_{\chi_v}(s,\pi_v)=C_{\chi_v}(s,\pi_v)\lambda_{\chi_v}(-s,w_0(\pi_v))\cdot A(s,\pi_v,w_0). \tag{3.3}$$

This is what we call the *Local Coefficient* attached to s,π_v,χ_v and w_0. The choice of w_0 is now specified by our fixed splitting as in [43].

Finally, if

$$E_\chi(s,\phi_s,g,P)=\int_{\mathbf{U}(F)\backslash U}E(s,\phi_s,ug,P)\overline{\chi(u)}du \tag{3.4}$$

is the χ-nonconstant term of the Eisenstein series, then ([7,41,42])

$$E_\chi(s,\phi_s,e,P)=\prod_{v\in S}W_v(e)\prod_{i=1}^{m}L_S(1+is,\pi,\tilde{r}_i)^{-1}, \tag{3.5}$$

where now S is assumed to have the property that if $v\notin S$, then χ_v is also unramified.

Applying Definition (3.4) to both sides of (2.6), using (3.5) now implies the *crude functional equation* ([40,41])

$$\prod_{i=1}^{m}L_S(is,\pi,r_i)=\prod_{v\in S}C_{\overline{\chi}_v}(s,\tilde{\pi}_v)\prod_{i=1}^{m}L_S(1-is,\pi,\tilde{r}_i). \tag{3.6}$$

4. The main induction, functional equations and multiplicativity

To prove the functional equation for each r_i with precise root numbers and L-function, we use (cf. [42]):

Proposition 4.1. *Given $1<i\leq m$, there exists a quasisplit guoup $\mathbf{G}_i$ over F, a maximal F-parabolic subgroup $\mathbf{P}_i=\mathbf{M}_i\mathbf{N}_i$, both unramified for every $v\notin S$, and a cuspidal automorphic form π' of $M_i=\mathbf{M}_i(\mathbb{A}_F)$, unramified for every $v\notin S$, such that if the adjoint action r' of ${}^LM_i$ on ${}^L\mathfrak{n}_i$ decomposes as $r'=\bigoplus_{j=1}^{m'}r'_j$, then*

$$L_S(s,\pi,r_i)=L_S(s,\pi',r'_1).$$

Moreover $m'<m$.

Remark 4.2. As was observed by Arthur [1], each $\mathbf{M}_i$ can be taken equal to $\mathbf{M}$ and $\pi'=\pi$. In fact each $\mathbf{G}_i$ can be taken to be an endoscopic group for $\mathbf{G}$, sharing $\mathbf{M}$ as a Levi subgroup. We shall record this as

Proposition 4.3. *Given i, $1<i\leq m$, there exist a quasisplit connected reductive F-group with $\mathbf{M}$ as a Levi subgroup and $m'<m$ for which $r'_1=r_i$.*

Using this induction and local-global arguments (cf. Proposition 5.1 of [42]), it was proved in [42] that

Theorem 4.4. (Theorems 3.5 and 7.7 of [42]) *a) For each i, $1 \leq i \leq m$, and each v, there exist a local L-function $L(s, \pi_v, r_{i,v})$, which is the inverse of a polynomial in q_v^{-s} whose constant term is 1, if $v < \infty$, and is the Artin L-function attached to $r_i \cdot \varphi'_v$, where $\varphi'_v : W'_{F_v} \to {}^L M_v$ is the homomorphism of the Deligne-Weil group into ${}^L M_v$ parametrizing π_v, if either $v = \infty$ or π_v has an Iwahori-fixed vector; and a root number $\varepsilon(s, \pi_v, r_{i,v}, \varphi_v)$ satisfying the same provisions, such that if*

$$L(s, \pi, r_i) = \prod L(s, \pi_v, r_{i,v}) \tag{4.1}$$

and

$$\varepsilon(s, \pi, r_i) = \prod_v \varepsilon(s, \pi_v, r_{i,v}, \psi_v), \tag{4.2}$$

then

$$L(s, \pi, r_i) = \varepsilon(s, \pi, r_i) L(1 - s, \pi, \widetilde{r}_i). \tag{4.3}$$

b) Let

$$\gamma(s, \pi_v, r_{i,v}, \psi_v) = \varepsilon(s, \pi_v, r_{i,v}, \psi_v) L(1 - s, \pi_v, \widetilde{r}_{i,v}) / L(s, \pi_v, r_{i,v}). \tag{4.4}$$

Then each $\gamma(s, \pi_v, r_{iv}, \psi_v)$ is multiplicative in the sense of equation (3.13) in Theorem 3.5 of [42]. (See below.) If π_v is tempered, then $\gamma(s, \pi_v, r_{i,v}, \psi_v)$ determines the corresponding root number and L-function uniquely and in fact that is how they are defined. Suppose π_v is non-tempered, then each $L(s, \pi_v, r_{i,v})$ is determined by means of the analytic continuation of its quasi-tempered Langlands parameter and multiplicativity of corresponding γ-functions. More precisely, if σ_v is the quasitempered Langlands parameter that gives π_v as a subrepresentation, then

$$L(s, \pi_v, r_{i,v}) = \prod_{j \in S_i} L(s, \overline{w}_j(\sigma_v), r'_{i(j),v}), \tag{4.5}$$

where the notation is as in part 3) of Theorem 3.5 of [42], provided that every L-function on the right hand side is holomorphic for $\mathrm{Re}(s) > 0$, whenever σ_v is (unitary) tempered (Conjecture 7.1 of [42], proved in many cases [3.6.42]). The set $S_i, \overline{w}_j$ and $r'_{i(j)}$ are defined as follows in which we drop the index v. Assume $\pi \subset Ind_{M_\theta(N_\theta \cap M)\uparrow M} \sigma \otimes 1$, where $\mathbf{M}_\theta(\mathbf{N}_\theta \cap \mathbf{M})$ is a parabolic subgroup of $\mathbf{M}$ defined by a subset $\theta \subset \Delta$, the set of simple roots of $\mathbf{A}_0$. Let $\theta' = \widetilde{w}_0(\theta) \subset \Delta$ and fix a reduced decomposition $\widetilde{w}_0 = \widetilde{w}_{n-1} \cdots \widetilde{w}_1$ of $\widetilde{w}_0$ (Lemma 2.1.1 of [41]). For each j, there exists a unique root $\alpha_j \in \Delta$ such that $\widetilde{w}_j(\alpha_j) < 0$. For each j, $2 \leq j \leq n-1$, let $\overline{w}_j = \widetilde{w}_{j-1} \cdots \widetilde{w}_1$. Set $\overline{w}_1 = 1$. Let $\Omega_j = \theta_j \cup \{\alpha_j\}$, where $\theta_1 = \theta$, $\theta_n = \theta'$, and $\theta_{j+1} = \widetilde{w}_j(\theta_j)$, $1 \leq j \leq n-1$. Then $\mathbf{M}_{\Omega_j}$ contains $\mathbf{M}_{\theta_j}(\mathbf{N}_{\theta_j} \cap \mathbf{M}_{\Omega_j})$ as a maximal parabolic subgroup and $\overline{w}_j(\sigma)$ is a representation of M_{θ_j}. The L-group ${}^L M_\theta$ acts on the space of r_i, but no longer necessarily irreducibly. Given an irreducible constituent of this action, there exists a unique j, $1 \leq j \leq n-1$, which under $\overline{w}_j$ is equivalent to an irreducible constituent of the action of ${}^L M_{\theta_j}$ on the Lie algebra of the L-group of $\mathbf{N}_{\theta_j} \cap \mathbf{M}_{\Omega_j}$. Let $i(j)$ be the index of this subspace and denote by

$r'_{i(j)}$ the action of ${}^L M_{\theta_j}$ on it. Finally, let S_i denote the set of all such j's for a given i. (See Theorem 3.5 and Section 7 of [42]. Also see the discussion just before Proposition 5.2 of [2.8].)

Remark 4.5. If $\mathbf{G} = GL_{t+n}, \mathbf{M} = GL_t \times GL_n$ and $\pi = \otimes_v \pi_v$ and $\pi' = \otimes_v \pi'_v$ are cuspidal representations of $GL_t(\mathbb{A}_F$ and $GL_n(\mathbb{A}_F)$, then $m = 1$ and $L(s, \pi \otimes \widetilde{\pi}', r_1)$ is precisely the Rankin-Selberg product L-function $L(s, \pi \times \pi')$ attached to (π, π') (cf. [21,43,44]). In this case each of the local L-functions and root numbers are precisely those of Artin through parametrization which is now available for $GL_N(F_v)$ for any N due to the work Harris-Taylor [18] and Henniart [19]. As we explain later, this will also be the case for many of our local factors as a result of our new cases of functoriality which we shall soon explain. This is quite remarkable, since our factors are defined using harmonic analysis, as opposed to the very arithmetic nature of the definition given for Artin factors. This is a perfect example of how deep Langlands' conjectures are.

Remark 4.6. The multiplicativity of local factors, in the sense of Theorem 3.4, are absolutely crucial in establishing our new cases of functoriality throughout our proofs [12,23,28]. In fact, not only do we need them to prove our strong transfers, they are also absolutely necessary in establishing our weak ones.

5. Twists by highly ramified characters, holomorphy and boundedness

While the functional equations developed from our method are in perfect shape and completely general, nothing that general can be said about the holomorphy and possible poles of these L-functions. On the other hand, there has recently been some remarkable new progress on the question of holomorphy of these L-function, mainly due to Kim [24,25,31]. They rely on reducing the existence of the poles to that of existence of certain *unitary* automorphic forms, which in turn points to the existence of certain local unitary representations. One then disposes of these representations, and therefore the pole, by checking the corresponding unitary dual of the local group. In view of the functional equation, this needs to be checked only for $\mathrm{Re}(s) \geq 1/2$. In fact, to carry this out, one needs to verify that:

$$\begin{aligned} &\textit{Certain local normalized (as in [41]) intertwining operators} \\ &\textit{are holomorphic and non-zero for } \mathrm{Re}(s) \geq 1/2, \end{aligned} \tag{5.1}$$

in each case [24,25,31]. The main issue is that one cannot always get such a contradiction and rule out the pole. In fact, there are many unitary duals whose complementary series extend all the way to $\mathrm{Re}(s)=1$.

On the other hand, if one considers a highly ramified twist π_η (see Theorem 5.1 below) of π, then it can be shown quite generally that every $L(s, \pi_\eta, r_i)$ is entire (cf. [45] for its local analogue). In fact, if η is highly ramified, then $w_0(\pi_\eta) \ncong \pi_\eta$, whose negation is a necessary condition for $M(s, \pi_\eta)$ to have a pole, a basic fact from Langlands spectral theory of Eisenstein series (Lemma 7.5 of [33]). This was used by Kim [24], and in view of the present powerful converse theorems [8,9], that

is all one needs to prove our cases of functoriality [12,23,28,30]. To formalize this, we borrow the following proposition (Proposition 2.1) from [28], in order to state the result. It is a consequence of our general induction (Propositions 4.1 and 4.3) and [24].

Theorem 5.1. *Assume* (5.1) *is valid. Then there exists a rational character* $\xi \in X(\mathbf{M}_F)$ *with the following property: Let* S *be a non-empty finite set of finite places of* F. *For every globally generic cuspidal representation* π *of* $M = \mathbf{M}(\mathbb{A}_F)$, *there exist non-negative integers* f_v, $v \in S$, *depending only on the local central characters of* π_v *for all* $v \in S$, *such that for every grössencharacter* $\eta = \otimes_v \eta_v$ *of* F *for which conductor of* η_v, $v \in S$, *is larger than or equal to* f_v, *every* L*-function* $L(s, \pi_\eta, r_i)$, $i = 1, \cdots, m$, *is entire, where* $\pi_\eta = \pi \otimes (\eta \cdot \xi)$. *The rational character* ξ *can be simply taken to be* $\xi(m) = \det(Ad(m)|\mathfrak{n})$, $m \in \mathbf{M}$, *where* $\mathfrak{n}$ *is the Lie algebra of* $\mathbf{N}$.

The last ingredient in applying converse theorems is that of boundedness of each $L(s, \pi, r_i)$ in every vertical strip of finite width, away from its poles, which are finite in number, again using the functional equation and under Assumption (5.1). This was proved in full generality by Gelbart-Shahidi [15], using the theory of Eisenstein series via [33] and [36]. The main theorem of [15] (Theorem 4.1) is in full generality, allowing poles for L-functions. Here we will state the version which applies to our π_η.

Theorem 5.2. *Under Assumption* (5.1), *let* ξ *and* η *be as in Theorem 5.1. Assume* η *is ramified enough so that each* $L(s, \pi_\eta, r_i)$ *is entire. Then, given a finite real interval* I, *each* $L(s, \pi_\eta, r_i)$ *remains bounded for all* s *with* $\mathrm{Re}(s) \in I$.

The main difficulty in proving Theorem 5.2 is having to deal with reciprocals of each $L(s, \pi, r_i)$, $2 \leq i \leq m$, near and on the line $\mathrm{Re}(s)=1$, the edge of the critical strip, whenever $m \geq 2$, which is unfortunately the case for each of our cases of functoriality. We handle this by appealing to equations (3.5) and estimating the non-constant term (3.4) by means of [33,36].

6. New cases of functoriality

Langlands functoriality predicts that every homomorphism between L-groups of two reductive groups over a number field, leads to a canonical correspondence between automorphic representations of the two groups. The following instances of functoriality are quite striking and are consequences of applying recent ingenious converse theorems of Cogdell and Piatetski-Shapiro [8,9] to certain classes of L-functions whose necessary properties are obtained mainly from our method. (See [20] for an insightful survey.) We refer to [11] for more discussion of these results and the transfer from $GL_2(\mathbb{A}_F) \times GL_2(\mathbb{A}_F)$ to $GL_4(\mathbb{A}_F)$, using Rankin-Selberg method by Ramakrishnan [37]. (See [23] for a proof using our method.)

6.a. Let $\pi_1 = \otimes_v \pi_{1v}$ and $\pi_2 = \otimes_v \pi_{2v}$ be cuspidal representations of $GL_2(\mathbb{A}_F)$ and $GL_3(\mathbb{A}_F)$, respectively. For each v, let ρ_{iv} be the homomorphism of Deligne-Weil group into $GL_{i+1}(\mathbb{C})$, parametrizing π_{iv}, $i = 1, 2$. Let $\pi_{1v} \boxtimes \pi_{2v}$ be the irreducible admissible representation of $GL_6(F_v)$ attached to $\rho_{1v} \otimes \rho_{2v}$ via [18,19]. Set $\pi_1 \boxtimes \pi_2 = \otimes_v (\pi_{1v} \boxtimes \pi_{2v})$, an irreducible admissible representation of $GL_6(\mathbb{A}_F)$.

Next, let $\pi = \pi_1$, $\pi_v = \pi_{1v}$ and $\rho_v = \rho_{1v}$. Let $\mathrm{Sym}^3(\pi_v)$ be the irreducible admissible representation of $GL_4(F_v)$ attached to $\mathrm{Sym}^3(\rho_v)$ and set $\mathrm{Sym}^3(\pi) = \otimes_v \mathrm{Sym}^3(\pi_v)$, an irreducible admissible representation of $GL_4(\mathbb{A}_F)$. We have:

Theorem 6.1 [28,30]. *a) The representations $\pi_1 \boxtimes \pi_2$ and $Sym^3(\pi)$ are automorphic.*

b) $Sym^3(\pi)$ is cuspidal, unless π is either of dihedral or of tetrahedral type.

In view of [9], one needs to show that $L(s, (\pi_1 \boxtimes \pi_2) \times (\sigma \otimes \eta))$ is **nice** in the sense that it satisfies the contentions of Theorems 4.4.a, 5.1 and 5.2 for a highly ramified grössencharacter η, where σ is a cuspidal representation of $GL_n(\mathbb{A}_F)$, $n = 1, 2, 3, 4$, which is unramified in every place v where either π_{1v} or π_{2v} is ramified. In particular for each v, one of π_{1v}, π_{2v} or σ_v is in the principal series. It then follows from multiplicativity (cf. Theorem 4.4) and the main results of [43,44], that these L-functions are equal to certain L-functions defined from our method. More precisely, we can take $(\mathbf{G}, \mathbf{M})$ to be: a) $\mathbf{G} = SL_5$, $\mathbf{M}_D = SL_2 \times SL_3$; b) $\mathbf{G} = \mathrm{Spin}(10)$, $\mathbf{M}_D = SL_3 \times SL_2 \times SL_2$; c) $\mathbf{G} = E_6^{sc}$, $\mathbf{M}_D = SL_3 \times SL_2 \times SL_3$; d) $\mathbf{G} = E_7^{sc}$, $\mathbf{M}_D = SL_3 \times SL_2 \times SL_4$, according as $n = 1, 2, 3, 4$, respectively. This leads to a proof that $\pi_1 \boxtimes \pi_2$ is weakly automorphic. The strong transfer requires a lot more work, involving base change, both normal [2] and non-normal [22], and finally a local result [5]. Automorphy of $\mathrm{Sym}^3(\pi)$ is a consequence of applying the first part to $(\pi, \mathrm{Ad}(\pi))$, where $\mathrm{Ad}(\pi)$ is the adjoint of π, established by Gelbart-Jacquet [14]. It does not require the use of [5].

Observe that we have in fact proved that the homomorphisms $GL_2(\mathbb{C}) \otimes GL_3(\mathbb{C}) \subset GL_6(\mathbb{C})$ and Sym^3: $GL_2(\mathbb{C}) \to GL_4(\mathbb{C})$ are functorial. Neither are endoscopic.

6.b. Let $\Pi = \otimes_v \Pi_v$ be a cuspidal representation of $GL_4(\mathbb{A}_F)$ and let $\Lambda^2 : GL_4(\mathbb{C}) \to GL_6(\mathbb{C})$ be the exterior square map. Also with π as in 6.a, let $\mathrm{Sym}^4(\pi) = \otimes_v \mathrm{Sym}^4(\pi_v)$, where $\mathrm{Sym}^4(\pi_v)$ is attached to $\mathrm{Sym}^4(\rho_v)$. Then

Theorem 6.2 (cf. [23]). *a) The map Λ^2 is weakly functorial, in the sense that there exists an automorphic form on $GL_6(\mathbb{A}_F)$ whose local components are equal to $\Lambda^2(\Pi_v)$ for all v, except if $v|2$ or $v|3$. Here $\Lambda^2(\Pi_v)$ is defined by the local Langlands conjecture [18,19].*

b) $Sym^4(\pi)$ is an automorphic representation of $GL_5(\mathbb{A}_F)$.

We point out that b) is obtained by applying a) to $\mathrm{Sym}^3(\pi)$. a) is proved by applying our method to Spin groups (Case $D_n - 1$ of [40], $n = k + 4$, $k = 0, 1, 2, 3$).

Proposition 6.3 (cf. [29]). *$Sym^4(\pi)$ is cuspidal unless π is either of dihedral, tetrahedral or octahedral type.*

Let $\pi = \otimes_v \pi_v$ be a cusp form on $GL_2(\mathbb{A}_F)$. For each unramified v, let α_v and β_v be the Hecke eigenvalues of π_v. Then as corollary to Proposition 6.3 we have the following striking improvements towards Ramanujan and Selberg conjectures.

Corollary 6.4. *a) (cf. [29]) Assume F is an arbitrary number field. Then $q_v^{-1/9} < |\alpha_v|$ and $|\beta_v| < q_v^{1/9}$. b) (cf. [27]). Assume $F = \mathbb{Q}$. Then $p^{-7/64} \leq |\alpha_p|$ and $|\beta_p| \leq p^{7/64}$. Similar estimates are valid for the Selberg conjecture. More precisely, the smallest positive eigenvalue $\lambda_1(\Gamma)$ of the Laplace operator on $L^2(\Gamma \backslash \mathfrak{n})$ for every congruence subgroup Γ satisfies $\lambda_1(\Gamma) \geq \dfrac{975}{4096} \cong 0.2380\cdots$*

6.c. Let $i : Sp_{2n}(\mathbb{C}) \hookrightarrow GL_{2n}(\mathbb{C})$ be the natural embedding. Let $\pi = \otimes_v \pi_v$

be a generic cuspidal representation of $SO_{2n+1}(\mathbb{A}_F)$. For each unramified v, let $\{A_v\} \subset Sp_{2n}(\mathbb{C})$ be the Hecke-Frobenius conjugacy class parametrizing π_v. Let $i(\pi_v)$ be the unramified representation of $GL_{2n}(F_v)$ attached to $\{i(A_v)\}$. Then the main theorem of [12] proves:

Theorem 6.5 [12]. *The embedding i is weakly functorial, i.e., there exist an automorphic representation of $GL_{2n}(\mathbb{A}_F)$ whose components are equal to $i(\pi_v)$ for almost all v.*

This is proved by applying our method to maximal parabolics of appropriate odd special orthogonal groups (Case B_n of [40]). The strong transfer is now also established by Ginzburg-Rallis-Soudry [16] as well as Kim [26] by building upon Theorem 6.5.

Final Comments. Many other cases are in progress. Among them are a proof of the existence of Asai transfer [32] using our method, which was originally proved by Ramakrishnan [38], using the Rankin-Selberg method. This is the first case where one needs to use quasisplit groups. Since the issue of stability of root numbers [10] (cf. [11]) seems to be close to being settled by means of our method [46], many others transfers should now be available. A similar approach for nongeneric representations was initiated in [13].

References

[1] J. Arthur, *Endoscopic L-functions and a combinatorial indentity*, Dedicated to H.S.M. Coxeter, Canad. J. Math., **51** (1999), 1135–1148.

[2] J. Arthur and L. Clozel, *Simple algebras, Base change, and the Advanced Theory of the Trace Formula*, Annals of Math. Studies, no. 120, Princeton University Press, 1989.

[3] M. Asgari, *Local L-functions for split spinor groups*, Canad. J. Math., (to appear).

[4] A. Borel., *Automorphic L-functions*, Automorphic Forms and Automorphic Representations, Proc. Sympos. Pure Math., vol. 33; II, Amer. Math. Soc., Providence, RI, 1979, 27–61.

[5] C. J. Bushnell and G. Henniart, *On certain dyadic representations*, Annals of Math., Appendix to [28], (to appear).

[6] W. Casselman and F. Shahidi, *On irreducibility of standard modules for generic representations*, Ann. Scient. Éc.Norm.Sup., **31** (1998), 561–589.

[7] W. Casselman and J. A. Shalika, *The unramified principal series of p-adic groups II; The Whittaker function*, Comp. Math., **41** (1980), 207–231.

[8] J. W. Cogdell and I. I. Piatetski-Shapiro, *Converse theorems for GL_n*, Publ. Math. IHES, **79** (1994), 157–214.

[9] _____, *Converse theorems for GL_n II*, J. Reine Angew. Math., **507** (1999), 165–188.

[10] _____, *Stability of gamma factors for $SO(2n+1)$*, Manuscripta Math., **95** (1998), 437–461.

[11] _____, *Converse Theorems, Functoriality and Applications to Number Theory*, These Proceedings.

[12] J. W. Cogdell, H. Kim, I. I. Piatetski-Shapiro and F. Shahidi, *On lifting from classical groups to GL_N*, Publ. Math. IHES, **93** (2001), 5–30.
[13] S. Friedberg and D. Goldberg, *On local coefficients for nongeneric representations of some classical groups*, Comp. Math., **116** (1999), 133–166.
[14] S. Gelbart and H. Jacquet, *A relation between automorphic representations of $GL(2)$ and $GL(3)$*, Ann. Scient. Éc. Norm. Sup., **11**(1978), 471–552.
[15] S. Gelbart and F. Shahidi, *Boundedness of automorphic L-functions in vertical strips*, Journal of AMS, **14** (2001), 79–107.
[16] D. Ginzburg, S. Rallis and D. Soudry, *Generic automorphic forms on SO(2n+1): functorial lift to $GL(2n)$, endoscopy and base change*, IMRN **14** (2001), 729–764.
[17] Harish-Chandra, *Automorphic forms on semisimple Lie groups*, SLN **62** (1968), Berlin-Heidelberg-New York.
[18] M. Harris and R. Taylor, *On the geometry and cohomology of some simple Shimura varieties*, Annals of Math. Studies, no. 151, Princeton University Press, 2001.
[19] G. Henniart, *Une preuve simple des conjectures de Langlands pour GL(n) sur un corps p-adique*, Invefnt. Math., **139** (2000), 439–455.
[20] G. Henniart, *Progrès récents en fonctorialité de Langlands*, Seminaire Bourbaki, Juin 2001, Exposé 890, 890-1 to 890-21.
[21] H. Jacquet, I. Piatetski-Shapiro and J. Shalika, *Rankin-selberg convolutions*, Amer. J. Math., **105** (1983), 367–464.
[22] _____, *Relvement cubique non normal*, C. R. Acad. Sci. Paris Sr. I Math., **292**, no. 12 (1981), 567–571.
[23] H. Kim, *Functoriality for the exterior square of GL_4 and symmetric fourth of GL_2*, preprint (2000).
[24] _____, *Langlands-Shahidi method and poles of automorphic L-functions: Application to exterior square L-functions*, Can. J. Math., **51** (1999), 835–849.
[25] _____, *Langlands-Shahidi method and poles of automorphic L-functions II*, Israel J. Math., **117** (2000), 261–284.
[26] _____, *Residual spectrum of odd orthogonal groups*, IMRN, **17** (2000), 873–906.
[27] H. Kim and P. Sarnak, *Refined estimates towards the Ramanujan and Selberg conjectures*, Appendix 2 to [23].
[28] H. Kim and F. Shahidi, *Functorial products for $GL_2 \times GL_3$ and the symmetric cube for GL_2*, Annals of Math., (to appear).
[29] _____, *Cuspidality of symmetric powers with applications*, Duke Math. J., **112** (2002), 177–197.
[30] _____, *Functorial products for $GL_2 \times GL_3$ and functorial symmetric cube for GL_2*, C.R. Acad. Sci. Paris, **331** (2000), 599–604.
[31] _____, *Symmetric cube L-functions for GL_2 are entire*, Ann. of Math., **150** (1999), 645–662.
[32] M. Krishnamurthy, The weak Asai transfer to $GL(4)$ via Langlands-Shahidi method, Thesis, Purdue University (2002).
[33] R. P. Langlands, *On the Functional Equations Satisfied by Eisenstein Series*,

Lecture Notes in Math., Vol 544, Springer-Verlag, 1976.
[34] _____, *Euler Products*, Yale University Press, 1971.
[35] C. Moeglin and J.-L. Waldspurger, *Spectral decomposition and Eisenstein series*, Cambridge Tracts in Math., vol. 113, Cambridge University Press, 1995.
[36] W. Müller, *The trace class conjecture in the theory of automorphic forms*, Ann. of Math., **130** (1989), 473–529.
[37] D. Ramakrishnan, *Modularity of the Rankin-Selberg L-series, and multiplicity one for SL*(2), Ann. of Math., **152** (2000), 45–111.
[38] _____, *Modularity of solvable Artin representations of GO*(4)*-type*, IMRN, **1** (2002), 1–54.
[39] F. Shahidi, *Functional equation satisfied by certain L-functions*, Comp. Math., **37**(1978), 171–207.
[40] _____, *On the Ramanujan conjecture and finiteness of poles for certain L-functions*, Ann. of Math., **127** (1988), 547–584.
[41] _____, *On certain L-functions*, Amer, J. Math., **103**(1981), 297–355.
[42] _____, *A proof of Langlands conjecture on Plancherel measures; Complementary series for p-adic groups*, Annals of Math., **132** (1990), 273–330.
[43] _____, *Local coefficients as Artin factors for real groups*, Duke Math. J., **52** (1985), 973–1007.
[44] _____, *Fourier transforms of intertwining operators and Plancherel measures for GL*(n), Amer. J. of Math., **106** (1984), 67–111.
[45] _____, *Twists of a general class of L-functions by highly ramified characters*, Canad. Math. Bull., **43** (2000), 380–384.
[46] _____, *Local coefficients as Mellin transforms of Bessel functions; Towards a general stability*, preprint (2002).

ICM 2002 · Vol. II · 667–677

Modular Representations of p-adic Groups and of Affine Hecke Algebras

Marie-France Vignéras*

Abstract

I will survey some results in the theory of modular representations of a reductive p-adic group, in positive characteristic $\ell \neq p$ and $\ell = p$.

2000 Mathematics Subject Classification: 11S37, 11F70, 20C08, 20G05, 22E50.
Keywords and Phrases: Modular representation, Reductive p-adic group, Affine Hecke algebra.

Introduction The congruences between automorphic forms and their applications to number theory are a motivation to study the smooth representations of a reductive p-adic group G over an algebraically closed field R of *any characteristic.* The purpose of the talk is to give a survey of some aspects of the theory of R-representations of G. In positive characteristic, most results are due to the author; when proofs are available in the litterature (some of them are not !), references will be given.

A prominent role is played by the unipotent block which contains the trivial representation. There is a finite list of types, such that the irreducible representations of the unipotent block are characterized by the property that they contain a unique type of the list. The types define functors from the R-representations of G to the right modules over generalized affine Hecke algebras over R with different parameters; in positive characteristic ℓ, the parameters are 0 when $\ell = p$, and roots of unity when $\ell \neq p$.

In characteristic 0 or $\ell \neq p$, for a p-adic linear group, there is a Deligne-Langlands correspondence for irreducible representations; the irreducible in the unipotent block are annihilated by a canonical ideal J; the category of representations annihilated by J is Morita equivalent to the affine Schur algebra, and the unipotent block is annihilated by a finite power J^k.

* Institut de Mathématiques de Jussieu, Université de Paris 7, France. E-mail: vigneras@math.jussieu.fr

New phenomena appear when $\ell = p$, as the supersingular representations discovered by Barthel-Livne and classified by Ch. Breuil for $GL(2, \mathbf{Q}_p)$. The modules for the affine Hecke algebras of parameter 0 and over R of characteristic p, are more tractable than the R-representations of the group, using that the center Z of a $\mathbf{Z}[q]$-affine Hecke algebra H of parameter q is a finitely algebra and H is a generated Z-module. The classification of the simple modules of the pro-p-Iwahori Hecke algebra of $GL(2, F)$ suggests the possibility of a Deligne-Langlands correspondence in characteristic p.

Complex case

Notation. $\mathbf{C}$ is the field of complex numbers, $G = \underline{G}(F)$ is the group of rational points of a reductive connected group $\underline{G}$ over a local non archimedean field F with residual field of characteristic p and of finite order q, and $\mathrm{Mod}_{\mathbf{C}}\, G$ is the category of complex smooth representations of G. All representations of G will be smooth, the stabilizer of any vector is open in G. An abelian category C is right (left) Morita equivalent to a ring A when C is equivalent to the category of right (left) A-modules.

The modules over complex affine Hecke algebras with parameter q are related by the Borel theorem to the complex representations of reductive p-adic groups.

Borel Theorem *The unipotent block of* $\mathrm{Mod}_{\mathbf{C}}\, G$ *is (left and right) Morita equivalent to the complex Hecke algebra of the affine Weyl group of* G *with parameter* q.

The proof has three main steps, in reverse chronological order, Bernstein a) [B] [BK], Borel b) [Bo], [C], Iwahori-Matusmoto c) [IM], [M].

a) **(1.a.1)** $\mathrm{Mod}_{\mathbf{C}}\, G$ *is a product of indecomposable abelian subcategories "the blocks".*

The unipotent block contains the trivial representation. The representations in the unipotent block will be called unipotent, although this term is already used by Lusztig in a different sense.

(1.a.2) *The irreducible unipotent representations are the irreducible subquotients of the representations parabolically induced from the unramified characters of a minimal parabolic subgroup of* G.

b) Let I be an Iwahori subgroup of G (unique modulo conjugation).

(1.b.1) *The category of complex representations of* G *generated by their* I*-invariant vectors is abelian, equivalent by the functor*

$$V \mapsto V^I = \mathrm{Hom}_{\mathbf{C}G}(\mathbf{C}[I\backslash G], V)$$

to the category $\mathrm{Mod}\, H_{\mathbf{C}}(G, I)$ *of right modules of the Iwahori Hecke algebra*

$$H_{\mathbf{C}}(G, I) = \mathrm{End}_{\mathbf{C}G}\, \mathbf{C}[I\backslash G].$$

(1.b.2) *This abelian category is the unipotent block.*

c) **(1.c)** *The Iwahori Hecke algebra $H_{\mathbf{C}}(G, I)$ is the complex Hecke algebra of the affine Weyl group of G with parameter q.*

The algebra has a very useful description called the Bernstein decomposition [L1] [BK], basic for the geometric description of Kazhdan-Lusztig [KL].

From (1.b.1), the irreducible unipotent complex representations of G are in natural bijection with the simple modules of the complex Hecke algebra $H_{\mathbf{C}}(G, I)$. By the "unipotent" Deligne-Langlands correspondence, the simple $H_{\mathbf{C}}(G, I)$-modules "correspond" to the G'-conjugacy classes of pairs (s, N), where $s \in G'$ is semisimple, $N \in \mathrm{Lie}\, G'$ and $\mathrm{Ad}(s)N = qN$, where G' is the complex dual group of G with Lie algebra $\mathrm{Lie}\, G'$. This is known to be a bijection when $G = GL(n, F)$ [Z] [R]. When G is adjoint and unramified (quasi-split and split over a finite unramified extension), it is also known to be a bijection if one adds a third ingredient, a certain irreducible geometric representation ρ of the component group of the simultaneous centralizer of both s and N in G'; this is was done by Chriss [C], starting from the basic case where G is split of connected center treated by Kazhdan Lusztig [KL] and by Ginsburg [CG] *. The adjoint and unramified case is sufficient for many applications to automorphic forms; to my knowledge the general case has not been done.

According to R. Howe, the complex blocks should be parametrized by types. The basic type, the trivial representation of an Iwahori subgroup, is the type of the unipotent block. An arbitrary block should be right Morita equivalent to the Hecke algebra of the corresponding type. The Hecke algebra of the type should be a generalized affine Hecke complex algebra with different parameters equal to positive powers of p. This long program started in 1976 is expected to be completed soon. The most important results are those of Bushnell-Kutzko for $GL(n, F)$ [BK], of Morris for the description of the Hecke algebra of a type [M], of Moy and Prasad for the definition of unrefined types [MP].

Conjecturally, the classification of simple modules over complex generalized affine Hecke algebras and the theory of types will give the classification of the complex irreducible representations of the reductive p-adic groups.

We consider now the basic example, the general linear p-adic group $GL(n, F)$. The the complex irreducible representations of $GL(n, F)$ over R are related by the "semi-simple" Deligne-Langlands correspondence (proved by Harris-Taylor [HT1]

* Introduction page 18. Complex representations of the absolute Weil-Deligne group with semi-simple part trivial on the *inertia subgroup* (6.1) are in natural bijection with the ℓ-adic representations of the absolute Weil group trivial on the *wild ramification subgroup* for any prime number $\ell \neq p$ [T] [D]. In the Deligne-Langlands correspondence, one considers only the representations which are Frobenius semi-simple.

and Henniart [He]), to the representations of the Galois group $\mathrm{Gal}(\overline{F}/F)$ of a separable algebraic closure $\overline{F}$ of F.

Deligne-Langlands correspondence

(1.d) *The blocks of* $\mathrm{Mod}_{\mathbf{C}}\, GL(n,F)$ *are parametrized by the conjugacy classes of the semi-simple* n*-dimensional complex representations* τ *of the inertia group* $I(\overline{F}/F)$ *which extend to the Galois group* $\mathrm{Gal}(\overline{F}/F)$.

(1.e) *The block parametrized by* τ *is equivalent to the unipotent block of a product of linear groups* $G_\tau = GL(d_1,F_1)\times\ldots\times GL(d_r,F_r)$ *over unramified extensions* F_i *of* F *where* $\sum d_i[F_i:F]=n$.

(1.f) *The irreducible unipotent representations of* $GL(n,F)$ *are parametrized by the* $GL(n,R)$*-conjugacy classes of pairs* (s,N) *where* $s\in GL(n,\mathbf{C})$ *is semi-simple,* $N\in M(n,\mathbf{C})$ *is nilpotent, and* $sN=qNs$.

Modular case Let R be an algebraically closed field of any characteristic. When the characteristic of R is 0, the theory of representations of G is essentially like the complex theory, and the above results remain true although some proofs need to be modified and this is not always in the litterature. From now on, we will consider "modular or mod ℓ" representations, i.e. representations over R of characteristic $\ell>0$.

Banal primes Although a reductive p-adic group G is infinite, it behaves often as a finite group. Given a property of complex representations of G which has formally a meaning for mod ℓ representations of G, one can usually prove that outside a finite set of primes ℓ, the property remains valid. This set of primes is called "banal" for the given property.

For mod ℓ representations the Borel theorem is false, because the mod ℓ unipotent block of $GL(2,F)$ contains representations without Iwahori invariant vectors when $q\equiv -1 \bmod \ell$ [V1].

(2) *The Borel theorem is valid for* mod ℓ *representations when* ℓ *does not divide the pro-order of any open compact subgroup of* G.

These primes are banal for the three main steps in the proof of the complex theorem.

a) **(2.a)** *Any prime is banal for the decomposition of* $\mathrm{Mod}_R\, G$ *in blocks.*

The complex proof of (1.a.1) does not extend. There is a new proof relying of the theory of unrefined types [V5 III.6] when $\ell\neq p$.

b) (1.a.2), (1.b.1), (1.b.2) remain true because ℓ does not does not divide the pro-order of the Iwahori subgroup I and [V2] [V4]:

(2.b) *Any irreducible cuspidal* mod ℓ*-representation of* G *is injective and projective in the category of* mod ℓ*-representations of* G *with a given central character when* ℓ *is as in theorem 2.*

c) Any prime ℓ is banal for the Iwahori-Matsumoto step because the proofs of Iwahori-Matusmoto and of Morris are valid over $\mathbf{Z}$, and for any commutative ring A, the Iwahori Hecke A-algebra

$$H_A(G,I) = \mathrm{End}_{AG}\, A[I\backslash G] \simeq H_{\mathbf{Z}}(G,I) \otimes_{\mathbf{Z}} A$$

is isomorphic to the Hecke A-algebra of the affine Weyl group of G with parameter q_A where q_A is the natural image of q in A.

The primes ℓ of theorem 2 are often called *the banal primes of* G because such primes are banal for many properties. For example, the category of $\mathrm{mod}\,\ell$-representations of G with a given central character has finite cohomological dimension [V4]. In the basic example $GL(n,F)$, ℓ is banal when $\ell \neq p$ and the multiplicative order of q modulo ℓ is $> n$.

Limit primes The set of primes banal for (1.a.2), (1.b.1) is usually larger than the set of banal primes of G. The primes of this set which are not banal will be called, following Harris, the *limit primes* of G. In the basic example $GL(n,F)$, the limit primes ℓ satisfy $q \equiv 1 \bmod \ell$ and $\ell > n$ [V3]. For number theoretic reasons, the limit primes are quite important [DT] [Be] [HT2]. They satisfy almost all the properties of the banal primes. For linear groups, the limit primes are banal for the property that no cuspidal representation is a subquotient of a proper parabolically induced representation. This is may be true for G general.

Let $\overline{\mathbf{Q}}_\ell$ be an algebraic closure of the field $\mathbf{Q}_\ell$ of ℓ-adic numbers, $\overline{\mathbf{Z}}_\ell$ its ring of integers and $\overline{\mathbf{F}}_\ell$ its residue field. The following statements follow from the theory of types, or from the description of the center of the category of $\mathrm{mod}\,\ell$ representations (the Bernstein center).

(3.1) *The reduction gives a surjective map from the isomorphism classes of the irreducible cuspidal integral* $\overline{\mathbf{Q}}_\ell$*-representations of* G *to the irreducible cuspidal* $\overline{\mathbf{F}}_\ell$*-representations of* G*, when* ℓ *is a banal or a limit prime for* G.

Natural characteristic The interesting case where the characteristic of R is p is not yet understood. There is a simplification: R-representations of G have non zero vectors invariant by the pro-p-radical I_p of I. The irreducible are quotients of $R[I_p\backslash G]$.

Some calculations have been made for $GL(2,F)$ [BL] [Br] [V9]. A direct classification of the irreducible R-representations of $G = GL(2,\mathbf{Q}_p)$ [BL] [Br] and of the pro-p-Iwahori Hecke R-algebra $H_R(G,I_p) = \mathrm{End}_{RG}\, R[I_p\backslash G]$ (called a $\mathrm{mod}\,p$ pro-p-Iwahori Hecke algebra) shows:

(4.1) *Suppose* R *of characteristic* p*. The pro-*p*-Iwahori functor gives a bijection between the irreducible* R*-representations of* $GL(2,\mathbf{Q}_p)$ *and the simple right* $H_R(G,I_p)$*- modules.*

This is the "$\mathrm{mod}\,p$ simple Borel theorem" for the pro-p-Iwahori group of

$GL(2,\mathbf{Q}_p)$. In particular p is banal for the simple version of (1.b.1) when $G = GL(2,\mathbf{Q}_p)$. Irreducible mod p representations of $GL(2,F)$ which are non subquotients of parabolically induced representations from a character of the diagonal torus are called supersingular [BL]. There is a similar definition for the mod p simple supersingular modules of the pro-p-Iwahori Hecke algebra of $GL(2,F)$.

(4.2) *There is a natural bijection between the* mod p *simple supersingular modules of the* mod p *pro-p-Iwahori Hecke algebra of $GL(2,F)$ and the* mod p *irreducible dimension 2 representations of the absolute Weil group of F.*

This suggests the existence of a mod p Deligne-Langlands correspondence. Some computations are beeing made by R. Ollivier for $GL(3,F)$.

We end this section with a new result on affine Hecke algebras as in [L3], which is important for the theory of representations modulo p.

(4.3) *Let H be an affine Hecke $\mathbf{Z}[q]$-algebra of parameter q associated to a generalized affine Weyl group W. Then the center Z of H is a finitely generated $\mathbf{Z}[q]$-algebra and H is a finitely generated Z-module.*

The key is to prove that H has a $\mathbf{Z}[q]$-basis $(q^{k(w)}E_w)_{w\in W}$ where (E_w) is a Bernstein $\mathbf{Z}[q^{-1}]$-basis of $H[q^{-1}]$. The assertion (4.3) was known when the parameter q is invertible.

Non natural characteristic R an algebraically closed field of positive characteristic $\ell \neq p$. Any prime $\ell \neq p$ is banal for the "simple Borel theorem". The "simple Borel theorem" is true mod $\ell \neq p$.

(5.1) *Suppose $\ell \neq p$. The Iwahori-invariant functor gives a bijection between the irreducible R-representations of G with $V^I \neq 0$ and the simple right $H_R(G,I)$-modules.*

The existence of an Haar measure on G with values in R implies that $\mathrm{Mod}_R\, G$ is left Morita equivalent to the convolution algebra $H_R(G)$ of locally constant, compact distributions on G with values in R. When the pro-order of I is invertible in R, the Haar measure on G over R normalized by I is an idempotent of $H_R(G)$, and (5.1) could have been already proved by I. Schur [V3]. In general (5.1) follows from the fact that $R[I\backslash G]$ is "almost projective" [V5].

More generally, one expects that the Howe philosophy of types remains true for modular *irreducible* representations. Their classification should reduce to the classification of the simple modules for generalized affine Hecke R-algebras of parameters equal to 0 if $\ell = p$, and to roots of unity if $\ell \neq p$. This is known for linear groups if $\ell \neq p$ [V5] or in characteristic $\ell = p$ for $GL(2,F)$ [V9].

The unipotent block is described by a finite set S of modular types, the "unipotent types" [V7]. The set S contains the class of the basic type (I, id). In the banal or limit case, this is the only element of S. A unipotent type (P,τ) is the G-

conjugacy class of an irreducible R-representation of a parahoric subgroup P of G, trivial on the pro-p-radical P_p, cuspidal as a representation of P/P_p (the group of rational points of a finite reductive group over the residual field of F). The isomorphism class of the compactly induced representation $\mathrm{ind}_P^G \tau$ of G determines the G-conjugacy class of τ, and conversely. We have $\mathrm{ind}_I^G \mathrm{id} = R[I\backslash G]$.

(5.2) Theorem *Suppose $\ell \neq p$. There exists a finite set S of types, such that*

- $\mathrm{ind}_P^G \tau$ is unipotent for any $(P,\tau) \in S$,

- an irreducible unipotent R-representation V of G is a quotient of $\mathrm{ind}_P^G \tau$ for a unique $(P,\tau) \in S$, called the type of V,

- the map $V \mapsto \mathrm{Hom}_{RG}(\mathrm{ind}_P^G \tau, V)$ between the irreducible quotients of $\mathrm{ind}_P^G \tau$ and the right $H_R(G,\tau) = \mathrm{End}_{RG} \mathrm{ind}_P^G \tau$ modules is a bijection.

The set S has been explicitely described only when G is a linear group [V5]. In the example of $GL(2,F)$ and $q \equiv -1$ modulo ℓ, the set S has two elements, the basic class and the class of $(GL(2,O_F),\tau)$ where τ is the cuspidal representation of dimension $q-1$ contained in the reduction modulo ℓ of the Steinberg representation of the finite group $GL(2,\mathbf{F}_q)$.

The Hecke algebra $H_R(G,\tau)$ of the type (P,τ) could probably be described a generalized affine Hecke R-algebra with different parameters (complex case [M] [L2], modular case for a finite group [GHM]).

The linear group in the non natural characteristic We consider the basic example $G = GL(n,F)$ and R an algebraically closed field of positive characteristic $\ell \neq p$.

(6.1) *Any prime $\ell \neq p$ is banal for the Deligne-Langlands correspondence.*

This means that (1.d) (1.e) (1.f) remain true when $\mathbf{C}$ is replaced by R. The proof is done by constructing congruences between automorphic representations for unitary groups of compact type [V6].

The unipotent block is partially described by the *affine Schur algebra*

$$S_R(G,I) = \mathrm{End}_{RG} V, \quad V = \oplus_{P \supset I} \mathrm{ind}_P^G \mathrm{id},$$

which is the ring of endomorphisms of the direct sum of the representations of G compactly induced from the trivial representation of the parahoric subgroups P containing the Iwahori subgroup I. The functor of I-invariants gives an isomorphism from the endomorphism ring of the RG-module V to the endomorphism ring of the right $H_R(G,I)$-module V^I and the $(S_R(G,I), H_R(G,I))$ module V^I satisfies the double centralizer property [V8].

(6.2) $\mathrm{End}_{H_R(G,I)} V^I = S_R(G,I)$, $\mathrm{End}_{S_R(G,I)} V^I = H_R(G,I)$.

In the complex case, the affine Schur algebra $S_{\mathbf{C}}(G,I)$ is isomorphic to an algebra already defined R.M. Green [Gr]: A complex *affine quantum linear group*

$\hat{U}(gl(n,q))$ has a remarkable representation W of countable dimension such that the tensor space $W^{\otimes n}$ satisfies the double centralizer property

$$\mathrm{End}_{\hat{S}(n,q)} W^{\otimes n} = \hat{H}(n,q), \quad \mathrm{End}_{\hat{H}(n,q)} W^{\otimes n} = \hat{S}(n,q)$$

where $\hat{S}(n,q)$ is the image of the action of $\hat{U}(gl(n,q))$ in $W^{\otimes n}$. *The algebras $\hat{S}(n,q)$ and $\hat{H}(n,q)$ are respectively isomorphic to $S_{\mathbf{C}}(G,I)$ and $H_{\mathbf{C}}(G,I)$; the bimodules $W^{\otimes n}$ and V^I are isomorphic.*

Let J be the annihilator of $R[I\backslash G]$ in the global Hecke algebra $H_R(G)$.

(6.3) Theorem *Suppose $\ell \neq p$.*

There exists an integer $k > 0$ such that the unipotent block of $\mathrm{Mod}_R G$ *is the set of R-representations of G annihilated by J^k.*

An irreducible representation of G is unipotent if and only if it is a subquotient of $R[I\backslash G]$, if and only if it is annihilated by J.

The abelian subcategory of representations of G annihilated by J is Morita equivalent to the affine Schur algebra $S_R(G,I)$.

This generalizes the Borel theorem to mod ℓ representations when G is a linear group. The affine Schur algebra exists and the double centralizer property (6.2) is true for a general reductive p-adic group G; in the banal case, the affine Schur algebra is Morita equivalent to the affine Hecke algebra.

Integral structures Let ℓ be any prime number. There are two notions of integrality for an admissible $\overline{\mathbf{Q}}_\ell$-representation V of G, $\dim V^K < \infty$ for all open compact subgroups K of G, which coincide when $\ell \neq p$ [V3]. One says that V is integral if V contains a G-stable $\overline{\mathbf{Z}}_\ell$-submodule generated by a $\overline{\mathbf{Q}}_\ell$-basis of V, and V is locally integral if the $H_{\overline{\mathbf{Q}}_\ell}(G,K)$-module V^K is integral, i.e. contains a $H_{\overline{\mathbf{Z}}_\ell}(G,K)$-submodule $\overline{\mathbf{Z}}_\ell$-generated by a $\overline{\mathbf{Q}}_\ell$-basis of V^K, for all K.

When V is irreducible and integral, the action of the center Z of G on V, the central character, is integral, i.e. takes values in $\overline{\mathbf{Z}}_\ell$. The situation is similar for a simple integral $H_{\overline{\mathbf{Q}}_\ell}(G,I)$-module W. The central character is integral, i.e. its restriction to the center of $H_{\overline{\mathbf{Z}}_\ell}(G,I)$ takes values in $\overline{\mathbf{Z}}_\ell$.

(7.1) Théorème

a) An irreducible cuspidal $\overline{\mathbf{Q}}_\ell$-representation V of G is integral if and only if its central character is integral.

b) A simple $H_{\overline{\mathbf{Q}}_\ell}(G,I)$-module is integral if and only if its central character is integral.

c) An irreducible representation V of G with $V^I \neq 0$ is locally integral if and only if V^I is an integral $H_{\overline{\mathbf{Q}}_\ell}(G,I)$-module.

The assertion b) results from (4.3). For a) [V3]. For $\ell = p$, c) is due to J.-F. Dat, using its theory of ℓ-adic analysis [D].

A general irreducible $\overline{\mathbf{Q}}_\ell$-representation V of G is contained in a parabolically induced representation of an irreducible cuspidal representation W of a Levi subgroup of G. If W is integral then V is integral, but the converse is false when $\ell = p$. When $\ell \neq p$, the converse is proved for classical groups by Dat using results of Moeglin (there is a gap in the "proof" of the converse in [V3]).

(7.2) Brauer-Nesbitt principle [V3][V11] *When $\ell \neq p$, the integral structures L of an irreducible $\overline{\mathbf{Q}}_\ell$-representation of G are $\overline{\mathbf{Z}}_\ell G$-finitely generated (hence commensurable) and their reduction $L \otimes \overline{\mathbf{F}}_\ell$ are finite length $\overline{\mathbf{F}}_\ell$-representations of G with the same semi-simplication (modulo isomorphism).*

When $\ell = p$, this is false. An integral cuspidal irreducible $\overline{\mathbf{Q}}_p$-representation V of G embeds in $\overline{\mathbf{Q}}_p[\Gamma\backslash G]$, for any discrete co-compact-mod-center subgroup Γ of G, and has a natural integral structure with an admissible reduction [V10]. When the theory of types is known, V is induced from an open compact-mod-center subgroup, hence has an integral structure with a non admissible reduction, which is not commensurable with the first one.

References

[BL] Barthel L., Livne R., Modular representations of GL_2 of a local field: the ordinary unramified case, J. of Number Theory 55, 1995, 1–27. Irreducible modular representations of GL_2 of a local field, Duke Math. J. 75, 1994, 261–292.

[Be] Bellaïche Joël, Congruences endoscopiques et représentations galoisiennes, Thèse Orsay 2002.

[B] Bernstein J.N., Le "centre" de Bernstein. Dans J.N. Bernstein, P. Deligne, D. Kazhdan, M.-F.Vignéras, Représentations des groupes réductifs sur un corps local, Travaux en cours. Hermann Paris 1984.

[Bo] Borel Armand, Admissible representations of a semisimple group over a local field with vectors fixed under an Iwahori subgroup, Invent. Math. 35, (1976), 233–259.

[Br] Breuil Christophe, Sur quelques représentations modulaires et p-adiques de $GL(2, \mathbf{Q}_p)$ I, II, Preprints 2001.

[BK] Bushnell Colin, Kutzko Phillip, The admissible dual of $GL(N)$ via compact open subgroups, Annals of Math. Studies, Princeton Univeristy Press, 129 (1993). Smooth representations of reductive p-adic groups: Structure theory via types, Proc. London Math. Soc. 77, (1988), 582–634.

[Ca] Cartier Pierre, Representations of p-adic groups: a survey, Proc. of Symp. in pure math. AMS XXXIII, part 1, 1979, 111–156.

[C] Chriss, Neil A., The classification of representations of unramified Hecke algebras, Math. Nachr. 191 (1998), 19–58.

[CG] Chriss N., Ginzburg V., Representation theory and complex geometry, Birkhauser 1997.

[D] Dat Jean-Francois, Generalized tempered representations of p-adic groups, Preprint 2002.

[DT] Diamond Fred, Taylor Richard, Non-optimal levels of mod ℓ representations, Invent. math. 115, (1994), 435–462.

[GHM] Geck, Meinolf; Hiss, Gerhard; Malle, Gunter, Towards a classification of the irreducible representations in non-describing characteristic of a finite group of Lie type. Math. Z. 221 (1996), no. 3, 353–386.

[Gr] Green R.M., The affine q-Schur algebra, Journal of Algebra 215 (1999) 379–411.

[IM] Iwahori N., Matsumoto H., On some Bruhat decompositions and the structure of the Hecke rings of p-adic Chevalley groups, Publ. Math. I.H.E.S. 25 (1965), 5–48.

[HT1] Harris Michael, Taylor Richard, The geometry and cohomology of some simple Shimura varieties, Annals of mathematics studies 151 (2001).

[HT2] Harris Michael, Taylor Richard, Notes on p-adic uniformization and congruences, 2002.

[H] Henniart Guy, Une preuve simple des conjectures de Langlands pour GL_n sur un corps p-adique, Invent. mat. 139 (2000), 339–350.

[KL] Kazhdan D., Lusztig G., Proof of the Deligne-Langlands conjecture for Hecke algebras, Invent. Math. 87, (1987), 153–215 .

[L1] Lusztig G., Some examples of square integrable representations of p-adic groups, Trans. Amer. Math. Soc. 277, (1983), 623–653.

[L2] Lusztig G., Classification of unipotent representations of simple p-adic groups, International Mathematics Research Notices No11, (1995), 517–589 .

[L3] Lusztig G., Representations of affine Hecke algebras, Soc. Math. de France, Astérisque 171-171 (1989), 73–84.

[M] Morris Lawrence, Tamely ramified intertwining algebras, Invent. math. 114, (1993),1-54. Tamely ramified supercuspidal representations, Ann. Sci. cole Norm. Sup. (4) 29 no. 5, (1996), 639–667.

[MP] Moy, Allen; Prasad, Gopal. Jacquet functors and unrefined minimal K-types, Comment. Math. Helv. 71 (1996), no. 1, 98-121. Unrefined minimal K-types for p-adic groups. Invent. Math. 116 (1994), no. 1-3, 393–408.

[R] Rogawski John, On modules over the Hecke algebra of a p-adic group, Invent. math. 79, (1985), 443–465.

[V1] Vignéras M.-F., Représentations modulaires de $GL(2, F)$ en caractéristique l, F corps p-adique, $p \neq l$, Compositio Mathematica 72 (1989), 33–66. Erratum, Compositio Mathematica 101,(1996), 109–113.

[V2] Vignéras M.-F., Banal Characteristic for Reductive p-adic Groups, J. of Number Theory Vol.47, Number 3, 1994, 378–397.

[V3] Vignéras M.-F., Représentations l-modulaires d'un groupes réductif p-adique avec $l \neq p$, Birkhauser Progress in Math.137 (1996).

[V4] Vignéras M.-F., Cohomology of sheaves on the building and

R-representations, Inventiones Mathematicae 127, 1997, 349–373.

[V5] Vignéras M.-F., Induced representations of reductive p-adic groups in characteristic $l \neq p$, Selecta Mathematica New Series 4 (1998) 549–623.

[V6] Vignéras M.-F., Correspondance locale de Langlands semi-simple pour $GL(n, F)$ modulo $\ell \neq p$, Inventiones 144, 2001, 197–223.

[V7] Vignéras M.-F., Irreducible modular representations of a reductive p-adic group and simple modules for Hecke algebras, International European Congress Barcelone 2000. Birkhauser Progress in Math. 201, 117–133.

[V8] Vignéras M.-F., Schur algebra of reductive p-adic groups I, Institut de Mathématiques de Jussieu, prépublication 289, Mai 2001, To appear in Duke Math. Journal

[V9] Vignéras M.-F., Representations modulo p of the p-adic group $GL(2, F)$, Institut de Mathématiques de Jussieu, prépublication 30, septembre 2001.

[V10] Vignéras M.-F., Formal degree and existence of stable arithmetic lattices of cuspidal representations of p-adic reductive groups, Invent. Math. 98 no. 3, (1989), 549–563.

[V11] Vignéras M.-F., On highest Whittaker models and integral structures, Institut de Mathématiques de Jussieu, prépublication 308, septembre 2001.

[Z] Zelevinski A., Induced representations of reductive p-adic groups II, Ann. scient. Ecole Norm. Sup. tome 13, (1980), 165–210.

Section 8. Real and Complex Analysis

A. Eremenko: *Value Distribution and Potential Theory* 681

Juha Heinonen: *The Branch Set of a Quasiregular Mapping* 691

Carlos E. Kenig: *Harmonic Measure and "Locally Flat" Domains* 701

Nicolas Lerner: *Solving Pseudo-Differential Equations* 711

C. Thiele: *Singular Integrals Meet Modulation Invariance* 721

S. Zelditch: *Asymptotics of Polynomials and Eigenfunctions* 733

Xiangyu Zhou: *Some Results Related to Group Actions in Several Complex Variables* 743

ICM 2002 · Vol. II · 681–690

Value Distribution and Potential Theory*

A. Eremenko†

Abstract

We describe some results of value distribution theory of holomorphic curves and quasiregular maps, which are obtained using potential theory. Among the results discussed are: extensions of Picard's theorems to quasiregular maps between Riemannian manifolds, a version of the Second Main Theorem of Nevanlinna for curves in projective space and non-linear divisors, description of extremal functions in Nevanlinna theory and results related to Cartan's 1928 conjecture on holomorphic curves in the unit disc omitting hyperplanes.

2000 Mathematics Subject Classification: 30D35, 30C65.
Keywords and Phrases: Holomorphic curves, Quasiregular maps, Meromorphic functions.

1. Introduction

Classical value distribution theory studies the following question: Let f be a meromorphic function in the plane. What can one say about solutions of the equation $f(z) = a$ as a varies? The subject was originated in 1880-s with two theorems of Picard (Theorems 1 and 4 below). An important contribution was made by E. Borel in 1897, who gave an "elementary proof" of Theorem 1, which opened a way to many generalizations. Borel's result (Theorem 12 below) also gives an extension of Picard's theorem to holomorphic curves $\mathbf{C} \to \mathbf{P}^n$. In 1925, R. Nevanlinna (partially in cooperation with F. Nevanlinna) created what is called now the Nevanlinna Theory of meromorphic functions, which was subject of intensive research [5]. A good elementary introduction to the subject is [18]. Griffiths and King [16] extended Nevanlinna theory to non-degenerate holomorphic maps $f : \mathbf{C}^n \to Y$, where Y is a compact complex manifold of dimension n. In modern times the emphasis has shifted to two multi-dimensional generalizations: holomorphic curves in complex manifolds and quasiregular mappings between real Riemannian manifolds. This survey is restricted to a rather narrow topic: generalizations of

*Supported by NSF grant DMS 0100512 and by the Humboldt Foundation.
†Department of Mathematics, Purdue University, West Lafayette IN 47907, USA. E-mail: eremenko@math.purdue.edu

Picard's theorem that are obtained with potential-theoretic methods. Some other applications of potential theory to value distribution can be found in [14, 20, 27]. Recent accounts of other methods in the theory of holomorphic curves are [21, 29].

We begin with Picard's Little Theorem:

Theorem 1 *Every entire function which omits two values in* $\mathbf{C}$ *is constant.*

To prove this by contradiction, we suppose that f is a non-constant entire function which omits 0 and 1. Then $u_0 = \log|f|$ and $u_1 = \log|f-1|$ are non-constant harmonic functions in the plane satisfying

$$|u_0^+ - u_1^+| \le c, \quad u_0 \vee u_1 \ge -c, \tag{1.1}$$

where $\vee$ stands for the pointwise sup, $u^+ = u \vee 0$, and c is a constant. There are several ways to obtain a contradiction from (1.1). They are based on rescaling arguments that permit to remove the c terms in (1.1). To be specific, one can find sequences $z_k \in \mathbf{C}$, $r_k > 0$ and $A_k \to +\infty$ such that $A_k^{-1} u_j(z_k + r_k z) \to v_j(z)$, $k \to \infty$, $|z| < 1$, $j = 0, 1$, where v_j are harmonic functions satisfying

$$v_0^+ = v_1^+, \quad v_0 \vee v_1 \ge 0, \quad v_j(0) = 0, \tag{1.2}$$

and $v_j \not\equiv 0$. This gives a contradiction with the uniqueness theorem for harmonic functions. The idea to base a proof of Picard's theorem on (1.2) comes from the paper [13] (the main result of this paper is described in Section 3 below). Two versions of the rescaling argument (existence of appropriate z_k, r_k and A_k) are given in [7, 12] and [19], respectively. The second version has an advantage that it uses only one result from potential theory, Harnack's inequality. Thus Picard's theorem can be derived from two facts: Harnack's inequality and the uniqueness theorem for harmonic functions. This makes the argument suitable for generalizations.

2. Quasiregular maps of Riemannian manifolds

We recall that a non-constant continuous map f between regions in $\mathbf{R}^n$ is called K-quasiregular if it belongs to the Sobolev class $W^{1,n}_{\mathrm{loc}}$ (first generalized derivatives are locally L^n-summable), and in addition

$$\|f'\|^n \le K J_f \quad \text{almost everywhere}, \tag{2.1}$$

where J is the Jacobian determinant and $K \ge 1$ is a constant. The standard references are [24, 25]. If $n = 2$, every quasiregular map can be factored as $g \circ \phi$, where g is analytic and ϕ a quasiconformal homeomorphism. It follows that Picard's Theorems 1 and 4 (below) extend without any changes to quasiregular maps of surfaces. For the rest of this section we assume that $n \ge 3$, and that all manifolds are connected. The weak smoothness assumption $f \in W^{1,n}_{\mathrm{loc}}$ is important: if we require more smoothness, the maps satisfying (2.1) will be local homeomorphisms (and even global homeomorphisms if the domain is $\mathbf{R}^n$). A fundamental theorem of Reshetnyak says that all quasiregular maps are open and discrete, that is they

have topological properties similar to those of analytic functions of one complex variable. Several other results about analytic functions have non-trivial extension to quasiregular mappings. One of the striking results in this area is Rickman's generalization of Picard's theorem [25]:

Theorem 2 *A K-quasiregular map $\mathbf{R}^n \to \mathbf{R}^n$ can omit only a finite set of points whose cardinality has an upper bound in terms on n and K.*

Even more surprising is that when $n = 3$, the number of omitted values can indeed be arbitrarily large, as Rickman's example in [26] shows.

It turns out that the method of proving Picard's theorem outlined in Section 1, extends to the case of quasiregular maps. One has to use a non-linear version of potential theory in $\mathbf{R}^n$ which is related to quasiregular maps in the same way as logarithmic potential theory to analytic functions. This relation between quasiregular maps and potential theory was discovered by Reshetnyak. He singled out a class of functions (which are called now A-harmonic functions), that share many basic properties (such as the maximum principle and Harnack's inequality) with ordinary harmonic functions, and such that $u \circ f$ is A-harmonic whenever u is A-harmonic and f quasiregular. In particular, $\log|x-a|$ is A-harmonic on $\mathbf{R}^n \backslash \{a\}$, so if f omits the value a, then $\log|f-a|$ satisfies Harnack's inequality (with constants depending on K and n). If m values are omitted by f we can obtain relations, similar to (1.2),

$$v_1^+ = \ldots = v_m^+, \quad v_i \vee v_j \geq 0, \quad v_j(0) = 0, \tag{2.2}$$

for certain A-harmonic functions $v_j \not\equiv 0$, $j = 1, \ldots, m$. Rickman's example mentioned above shows that such relations (2.2) are indeed possible with any given $m > 1$, which is consistent with the known fact that A-harmonic functions do not have the uniqueness property. However, an upper bound for m can be deduced from (2.2) using Harnack's inequality. This gives a pure potential-theoretic proof of Rickman's theorem [12, 19]. Notice that this proof does not depend on the deep result that quasiregular maps are open and discrete. Lewis's paper [19] which uses nothing but Harnack's inequality opened a path for further generalizations of Rickman's theorem. The strongest result in this direction was obtained by Holopainen and Rickman [17]. For simplicity, we state it only in the special case of quasiregular maps whose domain is $\mathbf{R}^n$.

Theorem 3 *Let Y be an orientable Riemannian manifold of dimension n. If there exists a K-quasiregular map $\mathbf{R}^n \to Y$, then the number of ends of Y has an upper bound that depends only on K and n.*

A more general result, with $\mathbf{R}^n$ replaced by a Riemannian manifold subject to certain conditions, is contained in [17].

Notice that there are no restrictions on Y in this theorem. Conditions of Theorem 3 will be satisfied if Y is a compact manifold with finitely many points removed, so a K-quasiregular map from $\mathbf{R}^n$ to a compact n-dimensional manifold can omit at most $N(K, n)$ points.

Now we turn to the second theorem of Picard mentioned in the Introduction:

Theorem 4 *If there exists a non-constant holomorphic map $f : \mathbf{C} \to S$ from the complex plane to a compact Riemann surface S, then the genus of S is at most 1.*

First extensions of this result to quasiregular maps in dimension $n > 2$ were obtained by Gromov in 1981 [6, Ch. 6] who proved that the fundamental group of a compact manifold of dimension n which receives a quasiregular map from $\mathbf{R}^n$ cannot be too large. Gromov applied a geometric method, based on isoperimetric inequalities, which goes back to Ahlfors's approach in dimension 2. The strongest result in this direction is the following theorem from [31]: *If a compact manifold Y of dimension $n \geq 2$ receives a quasiregular map from $\mathbf{R}^n$, then the fundamental group of Y is virtually nilpotent and has polynomial growth of degree at most n.*

We notice that unlike this last result, Theorem 3 has nothing to do with the fundamental group of Y: removing a finite set from a compact manifold does not change its fundamental group. Recently, Bonk and Heinonen [2] applied potential-theoretic arguments, somewhat similar to those outlined above, to obtain new topological obstructions to the existence of quasiregular maps:

Theorem 5 *If Y is a compact manifold of dimension n which receives a K-quasiregular map from $\mathbf{R}^n$, then the dimension of the de Rham cohomology ring of Y is bounded by a constant that depends only on n and K.*

This result implies that for every $K > 1$ there exist simply connected compact manifolds Y such that there are no K-quasiregular maps $\mathbf{R}^n \to Y$. The question whether there exists a compact simply connected manifold Y such that there are no quasiregular maps $\mathbf{R}^n \to Y$ (with any K) remains open.

For a compact manifold Y, the natural objects to pull back via f are differential forms rather then functions. According to the "non-linear Hodge theory" [28], each cohomology class of Y can be represented by a p-harmonic form, which satisfies a non-linear elliptic PDE. Such forms and their pullbacks to $\mathbf{R}^n$ play a similar role to the A-harmonic functions above.

It is natural to conjecture that the theorem of Bonk–Heinonen remains valid if the requirement that Y is compact is dropped. Such a generalization would also imply Theorem 3.

3. Holomorphic curves in projective varieties

Here we return to the classical logarithmic potential theory, which allows more precise quantitative estimates.

Points in the complex projective space $\mathbf{P}^n$ are represented by their homogeneous coordinates $z = (z_0 : \ldots : z_n)$. Let $Y \subset \mathbf{P}^n$ be an arbitrary projective variety. We consider divisors D on Y which are the zero sets of homogeneous forms $P(z_0, \ldots, z_n)$ restricted to Y. The degree of D is defined as the homogeneous degree of P. Suppose that q of such divisors D_j of degrees d_j are given, and they satisfy the condition that for some integer $m < q - 1$ every $m + 1$ of these divisors on Y have empty intersection. We are going to study the distribution of preimages of divisors D_j under a holomorphic map $f : \mathbf{C} \to Y$ whose image is not contained

in $\cup D_j$. To such a map correspond $n+1$ entire functions without common zeros: $f = (f_0, \ldots, f_n)$. Thus we are interested in the distribution of zeros of entire functions $P_j \circ f = P_j(f_0, \ldots, f_n)$.

We introduce the subharmonic functions

$$u = \|f\| = \sqrt{|f_0|^2 + \ldots + |f_n|^2} \quad \text{and} \quad u_j = \log |P_j \circ f|/d_j.$$

The assumption on intersections of D_j easily implies that

$$|\bigvee_{j\in I} u_j - u| \leq c \quad \text{for every} \quad I \subset \{1, \ldots, q\}, \quad \text{such that} \quad \operatorname{card} I = m+1. \quad (3.1)$$

This relation is a generalization of (1.1). The rescaling procedure mentioned in Section 1 permits to remove the constant c in (3.1) and obtain subharmonic functions $v_1, \ldots, v_q$ and v in a disc which satisfy

$$\bigvee_{j\in I} v_j = v, \quad I \subset \{1, \ldots, q\}, \quad \operatorname{card} I = m+1, \quad (3.2)$$

and such that v is *not harmonic.*

If f omits $q = 2m+1$ divisors in Y, then all v_j in (3.2) will be harmonic (while v is not!) and it is easy to obtain a contradiction. Indeed, let $E_j = \{z : v_j(z) = v(z)\}$. Then (3.2) with $q = 2m+1$ implies that for some I of cardinality $m+1$, the intersection $\cap_{j\in I} E_j$ has positive area. It follows by the uniqueness theorem that all v_j for $j \in I$ are equal. Applying (3.2) with this I we obtain that $v = v_j$ for $j \in I$, so v is harmonic, which gives a contradiction. Thus we obtain the following generalization of Picard's theorem proved by V. Babets in 1983 for the case $Y = \mathbf{P}^n$, $m = n$, and under a stronger restriction on the intersection of divisors [7].

Theorem 6 *Let Y be a projective variety. If a holomorphic map $\mathbf{C} \to Y$ omits $2m+1$ divisors, such that the intersection of any $m+1$ of them is empty, then f is constant.*

Notice that dimension of Y is not mentioned in this theorem. A more careful analysis of (3.2) and more sophisticated rescaling techniques yield a quantitative result of the type of the Nevanlinna's Second Main Theorem. To state it, we recall the definitions of Nevanlinna theory. If μ is the Riesz measure of u, then the Nevanlinna characteristic can be defined as

$$T(r, f) = \int_0^r \mu(\{z : |z| \leq t\}) \frac{dt}{t} - \log \|f(0)\|.$$

Let $n(r, D_j)$ be the number of zeros (counting multiplicity) of the entire function $P_j(f_0, \ldots, f_n)$ in the disc $\{z : |z| \leq r\}$, and

$$N(r, D_j, f) = \int_0^r n(t, D_j) \frac{dt}{t}, \quad (3.3)$$

supposing for simplicity that $g_j(0) \neq 0, j = 1, \ldots, q$. The following version of the Second Main Theorem was conjectured by Shiffmann in 1978 and proved in [13]:

Theorem 7 *Let Y be a projective variety, and q divisors D_j of degrees d_j in Y satisfy the intersection condition of Theorem 6. Let $f : \mathbf{C} \to Y$ be a holomorphic map whose image is not contained in $\cup_j D_j$. Then*

$$(q-2m)T(r,f) \le \sum_{j=1}^{q} \frac{1}{d_j} N(r, D_j, f) + o(T(r,f)),$$

when $r \to \infty$ avoiding a set of finite logarithmic measure.

This theorem is stated in [13] only for the case $Y = \mathbf{P}^n$, $m = n$ but the same proof applies to the more general statement. When $m = n = 1$ we obtain a rough form of the Second Main Theorem of Nevanlinna; with worse error term, and more importantly, without the ramification term. A corollary from Theorem 7 is the defect relation:

$$\sum_j \delta(D_j, f) \le 2m, \quad \text{where} \quad \delta(D_j, f) = 1 - \limsup_{r\to\infty} \frac{N(r, D_j, f)}{d_j T(r,f)}. \tag{3.4}$$

The key result of potential theory used in the proof of Theorem 7 is of independent interest [11]:

Theorem 8 *Suppose that a finite set of subharmonic functions $\{w_j\}$ in a region in the plane has the property that the pointwise minima $w_i \wedge w_j$ are subharmonic for every pair. Then the pointwise minimum of all these functions is subharmonic.*

This is derived in turn from the following:

Theorem 9 *Let G_1, G_2, G_3 be three pairwise disjoint regions, and μ_1, μ_2, μ_3 their harmonic measures. Then there exist Borel sets $E_j \subset \partial G_j$ such that $\mu_j(E_j) = 1$, $j = 1, 2, 3$, and $E_1 \cap E_2 \cap E_3 = \emptyset$.*

For regions in $\mathbf{R}^2$ (the only case needed for theorems 7 and 8) this is easy to prove: just take E_j to be the set of accessible points from G_j and notice that at most two points can be accessible from all three regions [11]. It is interesting that Theorem 9 holds for regions in $\mathbf{R}^n$ for all n, but the proof of this (based on advanced stochastic analysis rather then potential theory) is very hard [30].

We notice that the number 2 in Picard's Theorem 1, as well as in Theorem 7, thus admits an interpretation which seems to be completely different from the common one: with our approach it has nothing to do with the Euler characteristic of the sphere or its canonical class, but comes from Theorem 9. Recently, Siu [29] gave a proof of a result similar to Theorem 7 (with $Y = \mathbf{P}^n$, $m = n$) using different arguments which are inspired by "Vojta's analogy" between Nevanlinna theory and Diophantine approximation. However Siu's proof gives a weaker estimate $em \approx 2.718m$ instead of $2m$ in (3.4), and his assumptions on the intersection of divisors are stronger than those in Theorem 7.

The constant $2m$ in (3.4) is best possible. Moreover, one can give a rather complete characterization of extremal holomorphic curves of finite lower order. We recall that the lower order of a holomorphic curve is

$$\lambda = \liminf_{r\to\infty} \frac{\log T(r,f)}{\log r}.$$

Theorem 10 [8] *Let $D_1, \dots, D_q$ be divisors and f a curve satisfying all the hypotheses of Theorem 7. Suppose in addition that f has finite lower order and that equality holds in the defect relation* (3.4). *Then*
(i) 2λ *is an integer, and* $\lambda \geq 1$,
(ii) $T(r,f) = r^\lambda \ell(r)$, *where* $\ell(r)$ *is a slowly varying function in the sense of Karamata:* $\ell(cr)/\ell(r) \to 1$, $r \to \infty$ *uniformly with respect to* $c \in [1,2]$,
(iii) All defects are rational: $\delta(D_j, f) = p_j/\lambda$, *where* p_j *are integers whose sum is* $2m\lambda$.

When $m = n = 1$, this result was conjectured by F. Nevanlinna [23]. After long efforts, mainly by A. Pfluger, A. Edrei, W. Fuchs and A. Weitsman, D. Drasin finally proved F. Nevanlinna's conjecture in [4]. The potential-theoretic method presented here permitted to give a simpler proof of Drasin's theorem, and then to generalize the result to arbitrary dimension, as well as to obtain a stronger result in dimension 1 which is discussed in the next section. The proof of Theorem 10, is based on the following result about subharmonic functions:

Theorem 11 *Suppose that $v, v_1, \dots, v_q$, $q \geq 2m+1$ are subharmonic functions in the plane, which satisfy* (3.2), *and in addition* $v(z) \leq |z|^\lambda, \quad z \in \mathbf{C}$, *and* $v(0) = 0$. *Then the function*

$$h = \sum_{j=1}^{q} v_j - 2mv$$

is subharmonic. If h is harmonic, then 2λ is an integer and

$$v(re^{it}) = c|r|^\lambda |\cos \lambda(t-\alpha)|,$$

where $c > 0$ and α is a real constant.

4. Functions with small ramification

We recall the definition of the ramification term in Nevanlinna theory. Suppose that the image $f(\mathbf{C})$ of a holomorphic curve $f : \mathbf{C} \to \mathbf{P}^n$ is not contained in any hyperplane. This means that $f_0, \dots, f_n$ in the homogeneous representation of f are linearly independent. Let $n_1(r,f)$ be the number of zeros in the disc $\{z : |z| \leq r\}$ of the Wronski determinant $W(f_0, \dots, f_n)$, and $N_1(r,f)$ the averaged counting function of these zeros as in (3.3). If $n = 1$, then n_1 counts the number of critical points of f. The Second Main Theorem of Cartan [18] says that for every holomorphic curve f whose image does not belong to a hyperplane, and every finite set of hyperplanes $\{a_1, \dots, a_q\}$ in general position, we have

$$(q - n - 1 + o(1))T(r,f) + N_1(r,f) \leq \sum_{j=1}^{q} N(r, f, a_j), \tag{4.1}$$

when $r \to \infty$ avoiding a set of finite measure. This implies the defect relation

$$\sum_{j=1}^{q} \delta(a_j, f) + \theta(f) \leq n+1, \quad \text{where} \quad \theta(f) = \limsup_{r\to\infty} \frac{N_1(r,f)}{T(r,f)},$$

and $\delta(a,f)$ was defined in (3.4). So, if $n=1$, and the sum of deficiencies equals 2, then $\theta(f)=0$. The work of F. Nevanlinna [23] mentioned in Section 3 actually suggests something stronger than he conjectured: that the weaker assumption $\theta(f)=0$ for functions of finite lower order implies all conclusions (i)-(iii) of Theorem 10. This stronger result was proved in [9]. It follows that for functions of finite lower order the conditions $\theta(f)=0$ and $\sum\delta(a,f)=2$ are in fact equivalent. There is some evidence that this result might have the following extension to holomorphic curves in $\mathbf{P}^n$:

Conjecture *Let f be a holomorphic curve of finite lower order, whose image is not contained in any hyperplane. If $N_1(r)=o(T(r,f)), r\to\infty$, then λ is a rational number and assertion (ii) of Theorem 10 holds.*

This is not known even under a stronger assumption that the sum of deficiencies is $n+1$.

5. Cartan's conjecture

According to a philosophical principle of Bloch and Valiron [1], to theorems about entire functions should correspond theorems about families of functions in the unit disc, in the same way as Landau's theorem corresponds to Picard's theorem. One can supplement Theorem 6 with an explicit estimate of derivative of a holomorphic map from the unit disc to projective space that omits $2m+1$ hypersurfaces satisfying the intersection condition of Theorem 6. To prove such generalization of Landau's theorem, one replaces the use of the uniqueness theorem for harmonic functions by the corresponding quantitative result as in [22].

In 1887 Borel proved an extension of Picard's theorem, from which Theorem 6 and many other similar results (see, for example, [15]) can be derived:

Theorem 12 (Borel) *If $f_1,\dots,f_p$ are entire functions without zeros, that satisfy*

$$f_1+f_2+\dots+f_p=0, \tag{5.1}$$

then there is a partition of the set $J=\{f_1,\dots,f_p\}$ into classes I, such that for every I, all functions in I are proportional and their sum is zero.

When $p=3$ it is equivalent to the Picard's Little Theorem. The question is what kind of normality criterion corresponds to Theorem 12 in the same way as Montel's criterion corresponds to Picard's theorem. The following conjecture was stated by H. Cartan in his thesis [3] (see also [18] for a comprehensive discussion of this conjecture).

Conjecture A *Let F be an infinite sequence of p-tuples $f=(f_1,\dots,f_p)$ of holomorphic functions in the unit disc, such that each f_j has no zeros, and (5.1) is satisfied.*

Then there exists an infinite subsequence F' of F and a partition of the set $J=\{1,\dots,p\}$ into classes I, such that for f in F' and every class I we have:
$(*)$ there exists $j\in I$ such that for every $i\in I$ the ratios f_i/f_j are uniformly bounded on compact subsets of the unit disc, and $\sum_{i\in I} f_i/f_j\to 0$ uniformly on compact subsets of the unit disc.

One obtains this statement by replacing "proportional" by "have bounded ratio" and "equals zero" by "tends to zero" in the conclusions of Borel's theorem. When $p = 3$, Conjecture A is equivalent to Montel's theorem.

Let us call a subset $I \subset J = \{1, \dots, p\}$ having the property $(*)$ a *C-class* of the sequence F'. Cartan proved in [3] that under the hypotheses of Conjecture A there exists an infinite subsequence F', such that either the whole set J constitutes a single C-class, or there are at least 2 disjoint C-classes in J. This result implies that Conjecture A is true for $p = 4$, which corresponds to holomorphic curves in $\mathbf{P}^2$ omitting four lines. Indeed, it follows from $(*)$ that each C-class contains at least two elements, so if there are two disjoint C-classes they have to be a partition of the set J of four elements. For $p \geq 5$, Cartan's result falls short of proving his conjecture because the union of the two C-classes whose existence is asserted might not coincide with the whole set $\{1, \dots, p\}$.

It turns out that Conjecture A is wrong as originally stated, beginning from $p = 5$ (that is in dimensions ≥ 3). A simple counterexample was constructed in [10]). Nevertheless a small modification of the statement is valid in dimension 3:

Conjecture B *Under the assumptions of Conjecture A its conclusions hold is the disc* $\{z : |z| < r_p\}$, *where* $r_p < 1$ *is a constant that depends only on* p.

This was proved in [10] when $p = 5$, that is for holomorphic curves in $\mathbf{P}^3$ omitting 5 planes.

References

[1] A. Bloch, La conception actuelle de la théorie des fonctions entiéres et méromorphes, *Ens. Math.*, 25 (1926) 83–103.

[2] M. Bonk and J. Heinonen, Quasiregular mappings and cohomology, *Acta math.*, 186 (2001) 219–238.

[3] H. Cartan, Sur les systèmes de fonctions holomorphes a variétés linéaires lacunaires, *Ann. Sci. École Norm. Supèr.*, 45 (1928) 255–346.

[4] D. Drasin, Proof of a conjecture of F. Nevanlinna concerning functions which have deficiency sum two, *Acta math.*, 158 (1987) 1–94.

[5] A.A. Gol'dberg and I.V. Ostrovskii, *Distribution of values of meromorphic functions*, Moscow, Nauka, 1970 (Russian).

[6] M. Gromov, *Metric structures for Riemannian and non-Riemannian spaces*, Birkhauser, Boston, 1999.

[7] A. Eremenko, A Picard type theorem for holomorphic curves, *Period. Math. Hung.*, 38 (1999) 39–42.

[8] A. Eremenko, Extremal holomorphic curves for defect relations, *J. d'Analyse math.*, vol. 74 (1998) 307–323.

[9] A. Eremenko, Meromorphic functions with small ramification, *Indiana Univ. Math. J.*, 42, 4 (1993), 1193–1218.

[10] A. Eremenko, Holomorphic curves omitting five planes in projective space, *Amer. J. Math.*, 118 (1996), 1141–1151.

[11] A. Eremenko, B. Fuglede and M. Sodin, On the Riesz charge of the lower envelope of delta-subharmonic functions. *Potential Analysis*, 1 (1992) 191–204.

[12] A. Eremenko and J. Lewis, Uniform limits of certain A-harmonic functions with applications to quasiregular mappings, *Ann. Acad. Sci. Fenn., Ser. A. I.*, 16 (1991) 361–375.

[13] A. Eremenko and M. Sodin, Value distribution of meromorphic functions and meromorphic curves from the point of view of potential theory. *St. Petersburg Math. J.*, 3 (1992), 109–136.

[14] A. Eremenko and M. Sodin, Proof of a conditional theorem of Littlewood on the distribution of values of entire functions, Engl. transl.: *Math USSR Izvestiya*, 30 (1988) 395–402.

[15] M. Green, Some Picard theorems for holomorphic maps to algebraic varieties. *Amer. J. Math.* 97 (1975), 43–75.

[16] Ph. Griffiths and J. King, Nevanlinna theory and holomorphic mappings between algebraic varieties, *Acta Math.*, 130 (1973) 145–220.

[17] I. Holopainen and S. Rickman, Ricci curvature, Harnack functions and Picard type theorems for quasiregular mappings, *Analysis and topology*, 315–326. World Sci. Publ., River Edge NJ, 1998.

[18] S. Lang, *Introduction to complex hyperbolic spaces*, Springer, Berlin, 1987.

[19] J. Lewis, Picard's theorem and Rickman's theorem by way of Harnack's inequality, *Proc. AMS*, 122 (1994) 199–206.

[20] J. Lewis, On a conditional theorem of Littlewood for quasiregular entire functions, *J. Anal. Math.*, 62 (1994) 169–198.

[21] Min Ru, *Nevanlinna theory and its relation to Diophantine approximation*, World Scientific, River Edge, NJ, 2001.

[22] N. Nadirashvili, On a generalization of Hadamard's three-circle theorem, *Vestnik Moskovskogo Universiteta, Mat.*, 31 (1976) 39–42.

[23] F. Nevanlinna, Über eine Klasse meromorpher Funktionen, *7-e congrès des mathématiciens scandinaves, Oslo 1929*, A. W. Broggers, Oslo, 1930.

[24] Yu. Reshetnyak, *Space mappings with bounded distortion*, AMS, Providence, RI, 1989.

[25] S. Rickman, *Quasiregular mappings*, Springer, NY, 1993.

[26] S. Rickman, An analogue of Picard's theorem for quasiregular mappings in dimension three, *Acta math.* 154 (1985) 195–242.

[27] A. Russakovskii and B. Shiffman, Value distribution for sequences of rational mappings and complex dynamics, *Indiana Univ. Math. J.*, 46 (1997) 897–932.

[28] C. Scott, L^p theory of differential forms on manifolds, *Trans. AMS*, 347 (1995), 2075–2096.

[29] Y.-T. Siu, Recent techniques in hyperbolicity problems, *Several complex variables (Berkeley CA, 1995-1996)*, 429–508, Cambridge UP, 1999.

[30] B. Tsirelson, Triple points: from non-Brownian filtrations to harmonic measures, *GAFA*, 7 (1997) 1096–1142.

[31] N. Varopoulos, L. Saloff-Coste and T. Coulhon, *Analysis and geometry on groups*, Cambridge UP, Cambridge, 1992.

ICM 2002 · Vol. II · 691–700

The Branch Set of a Quasiregular Mapping*

Juha Heinonen†

Abstract

We discuss the issue of branching in quasiregular mapping, and in particular the relation between branching and the problem of finding geometric parametrizations for topological manifolds. Other recent progress and open problems of a more function theoretic nature are also presented.

2000 Mathematics Subject Classification: 30C65, 57M12.
Keywords and Phrases: Quasiregular map, Bi-Lipschitz map, Branch set.

1. Branched coverings

A continuous mapping $f : X \to Y$ between topological spaces is said to be a *branched covering* if f is an open mapping and if for each $y \in Y$ the preimage $f^{-1}(y)$ is a discrete subset of X. The *branch set* B_f of f is the closed set of points in X where f does not define a local homeomorphism.

Nonconstant holomorphic functions between connected Riemann surfaces are examples of branched coverings. From the Weierstrassian (power series) point of view this property of holomorphic functions is almost immediate. It is a deeper fact, due to Riemann, that the same conclusion can be drawn from the mere definition of complex differentiability, or, equivalently, from the Cauchy-Riemann equations. Most of this article discusses the repercussions of this fact.

2. Quasiregular mappings

In a 1966 paper [27], Reshetnyak penned a definition for *mappings of bounded distortion* or, as they are more commonly called today, *quasiregular mappings.*

*Supported by NSF grant DMS 9970427. I thank Mario Bonk and Alex Eremenko for their criticism on earlier versions of this article. My warmest thanks go to Mario Bonk, Seppo Rickman, and Dennis Sullivan for collaboration, mentoring, and friendship.

†Department of Mathematics, University of Michigan, MI 48109, USA. E-mail: juha@math.lsa.umich.edu

These are nonconstant mappings $f : \Omega \to \mathbf{R}^n$ in the Sobolev space $W^{1,n}_{loc}(\Omega; \mathbf{R}^n)$, where $\Omega \subset \mathbf{R}^n$ is a domain and $n \geq 2$, satisfying the following requirement: there exists a constant $K \geq 1$ such that

$$|f'(x)|^n \leq K J_f(x) \tag{2.1}$$

for almost every $x \in \Omega$, where $|f'(x)|$ denotes the operator norm of the (formal) differential matrix $f'(x)$ with $J_f(x)$=det$f'(x)$ its Jacobian determinant. One also speaks about *K-quasiregular mappings* if the constant in (2.1) is to be emphasized.[1]

Requirement (2.1) had been used as the analytic definition for quasiconformal mappings since the 1930s, with varying degrees of smoothness conditions on f. Quasiconformal mappings are by definition quasiregular homeomorphisms, and Reshetnyak was the first to ask what information inequality (2.1) harbours *per se*. In a series of papers in 1966–69, Reshetnyak laid the analytic foundations for the theory of quasiregular mappings. The single deepest fact he discovered was that quasiregular mappings are branched coverings (as defined above). It is instructive to outline the main steps in the proof for this remarkable assertion, which akin to Riemann's result exerts significant topological information from purely analytic data. For the details, see, *e.g.*, [28], [29], [18].

To wit, let $f : \Omega \to \mathbf{R}^n$ be $K-$quasiregular. Fix $y \in \mathbf{R}^n$ and consider the preimage $Z = f^{-1}(y)$. One first shows that the function $u(x) = \log|f(x) - y|$ solves a quasilinear elliptic partial differential equation

$$-\mathrm{div}\mathcal{A}(x, \nabla u(x)) = 0, \qquad \mathcal{A}(x, \xi) \cdot \xi \simeq |\xi|^n, \tag{2.2}$$

in the open set $\Omega \setminus Z$ in the weak (distributional) sense. In general, $\mathcal{A}$ in (2.2) depends on f, but its ellipticity only on K and n. For holomorphic functions, *i.e.*, for $n = 2$ and $K = 1$, equation (2.2) reduces to the Laplace equation $-\mathrm{div}\nabla u = 0$.

Now $u(x)$ tends to $-\infty$ continuously as x tends to Z. Reshetnyak develops sufficient *nonlinear potential theory* to conclude that such *polar sets*, associated with equation (2.2), have Hausdorff dimension zero. It follows that Z is totally disconnected, *i.e.*, the mapping f is *light*. This is the purely analytic part of the proof. The next step is to show that nonconstant quasiregular mappings are *sense-preserving*. This part of the proof mixes analysis and topology. What remains is a purely topological fact that sense-preserving and light mappings between connected oriented manifolds are branched coverings.

Initially, Reshetnyak's theorem served as the basis for a higher dimensional function theory. In the 1980's, it was discovered by researchers in nonlinear elasticity. In the following, we shall discuss more recent, different types of applications.

3. The branch set

Branched coverings between surfaces behave locally like analytic functions according to a classical theorem of Stoïlow. By a theorem of Chernavskiĭ, for every

[1]The definition readily extends for mappings between connected oriented Riemannian n-manifolds.

$n \geq 2$, the branch set of a discrete and open mapping between n-manifolds has topological dimension at most $n-2$. For branched coverings between 3-manifolds, the branch set is either empty or has topological dimension 1 [24], but in dimensions $n \geq 5$ there are branched coverings between n-manifolds with branch set of dimension $n-4$, cf. Section 7.[2]

The branch set of a quasiregular mapping is a somewhat enigmatic object in dimensions $n \geq 3$. It can be very complicated, containing for example many wild Cantor sets of classical geometric topology [14], [15]. There is currently no theory available that would explain or describe the geometry of allowable branch sets, cf. Problems **2** and **4** in Section 7.

In the next three sections, we shall discuss the problem of finding bi-Lipschitz parametrizations for metric spaces. It will become clear only later how this problem is related to the branch set.

4. Bi-Lipschitz parametrization of spaces

A homeomorphism $f : X \to Y$ between metric spaces is *bi-Lipschitz* if there exists a constant $L \geq 1$ such that

$$L^{-1} d_X(a,b) \leq d_Y(f(a), f(b)) \leq L d_X(a,b)$$

for each pair of points $a, b \in X$. It appears to be a difficult problem to decide when a given a metric space X can be covered by open sets each of which is bi-Lipschitz homeomorphic to an open set in $\mathbf{R}^n$, $n \geq 2$. If this is the case, let us say, for brevity and with a slight abuse of language, that X is *locally bi-Lipschitz equivalent to* $\mathbf{R}^n$.

Now a separable metrizable space is a Lipschitz manifold (in the sense of charts) if and only if it admits a metric, compatible with the given topology, that makes the space locally bi-Lipschitz equivalent to $\mathbf{R}^n$ [22]. The problem here is different from characterizing Lipschitz manifolds among topological spaces, for the metric is given first, cf. [8], [39], [40], [41].

To get a grasp of the difficulty of the problem, consider the following example: *There exist finite 5-dimensional polyhedra that are homeomorphic to the standard 5-sphere* $\mathbf{S}^5$*, but not locally bi-Lipschitz equivalent to* $\mathbf{R}^5$. This observation of Siebenmann and Sullivan [38] is based on a deep result of Edwards [9], which asserts that the double suspension $\Sigma^2 H^3$ of a 3-dimensional homology sphere H^3, with nontrivial fundamental group, is homeomorphic to the standard sphere $\mathbf{S}^5$. (See also [6].) One can think of $X = \Sigma^2 H^3$ as a join $X = \mathbf{S}^1 * H^3$, and it is easy to check that the complement of the *suspension circle* $\mathbf{S}^1$ in X is not simply connected. Consequently, every homeomorphism $f : X \to \mathbf{S}^5$ must transfer $\mathbf{S}^1$ to a closed curve $\Gamma = f(\mathbf{S}^1)$ whose complement in $\mathbf{S}^5$ is not simply connected. A general position argument and Fubini's theorem imply that, in this case, the Hausdorff dimension of Γ must be at least 3. Hence f cannot be Lipschitz. In fact, f cannot be Hölder continuous with any exponent greater than $1/3$. It is not known what other obstructions there are for a homeomorphism $X \to \mathbf{S}^5$, cf. [16, Questions 12–14].

See [33] and [37] for surveys on parametrization and related topics.

[2] See [23] for a recent survey on dimension theory and branched coverings.

5. Necessary conditions

What are the obvious necessary conditions that a given metric space X must satisfy, if it were to be locally bi-Lipschitz equivalent to $\mathbf{R}^n$, $n \geq 2$? Clearly, X must be an n-manifold. Next, bi-Lipschitz mappings preserve Hausdorff measure in a quantitative manner, so in particular X must be *n-rectifiable* in the sense of geometric measure theory; moreover, locally the Hausdorff n-measure should assign to each ball of radius $r > 0$ in X a mass comparable to r^n. Let us say that X is *metrically n-dimensional* if it satisfies these geometric measure theoretic requirements.

It is not difficult to find examples of metrically n-dimensional manifolds that are not locally bi-Lipschitz equivalent to $\mathbf{R}^n$. The measure theory allows for cusps and folds that are not tolerated by bi-Lipschitz parametrizations. Further geometric constraints are necessary; but, unlike in the case of the measure theoretic conditions, it is not obvious what these constraints should be. A convenient choice is that of *local linear contractibility*: locally each metric ball in X can be contracted to a point inside a ball with the same center but radius multiplied by a fixed factor. [3]

Still, a metrically n-dimensional and locally linearly contractible metric n-manifold need not be locally bi-Lipschitz equivalent to $\mathbf{R}^n$. The double suspension of a homology 3-sphere with nontrivial fundamental group as described in the previous section serves as a counterexample. In 1996, Semmes [34], [35] exhibited examples to the same effect in all dimensions $n \geq 3$, and recently Laakso [21] crushed the last hope that the above conditions might characterize at least 2-dimensional metric manifolds that are locally bi-Lipschitz equivalent to $\mathbf{R}^2$. However, unlike the examples of Edwards and Semmes, Laakso's metric space cannot be embedded bi-Lipschitzly in any finite dimensional Euclidean space. Thus the following problem remains open:

Problem 1 Let X be a topological surface inside some $\mathbf{R}^N$ with the inherited metric. Assume that X is metrically 2-dimensional and locally linearly contractible. Is X then locally by-Lipschitz equivalent to $\mathbf{R}^2$?

In conclusion, perhaps excepting the dimension $n = 2$, more necessary conditions are needed in order to characterize the spaces that are locally bi-Lipschitz equivalent to $\mathbf{R}^n$.[4] The idea to use Reshetnyak's theorem in this connection originates in two papers by Sullivan [40], [41], and is later developed in [17]. Recall that in this theorem topological conclusions are drawn from purely analytic data. Now imagine that such data would make sense in a space that is not *a priori* Euclidean. Then, if one could obtain a branched covering mapping into $\mathbf{R}^n$, *manifold points would appear, at least outside the branch set.* We discuss the possibility to develop this idea in the next section.

6. Cartan-Whitney presentations

[3] See [36] for analytic implications of this condition.

[4] There are interesting and nontrivial sufficient conditions known, but these are far from being necessary [42], [43], [2], [3], [5].

Let X be a metrically n-dimensional, linearly locally contractible n-manifold that is also a metric subspace of some $\mathbf{R}^N$. Suppose that there exists a bi-Lipschitz homeomorphism $f : X \to f(X) \subset \mathbf{R}^n$. Then f pulls back to X the standard coframe of $\mathbf{R}^n$, providing almost everywhere defined (essentially) bounded differential 1-forms $\rho_i = f^* dx_i$, $i = 1, \ldots, n$. To be more precise here, by Kirzsbraun's theorem, f can be extended to a Lipschitz mapping $\bar{f} : \mathbf{R}^N \to \mathbf{R}^n$, and the 1-forms

$$\rho_i = \bar{f}^* dx_i = d\bar{f}_i, \qquad i = 1, \ldots, n, \tag{6.1}$$

are well defined in $\mathbf{R}^N$ as flat 1-forms of Whitney. *Flat forms* are forms with L^∞-coefficients such that the distributional exterior differential of the form also has L^∞-coefficients. The forms in (6.1) are closed, because the fundamental relation $d\bar{f}^* = \bar{f}^* d$ holds true for Lipschitz maps.

According to a theorem of Whitney [45, Chapter IX], flat forms (ρ_i) have a well defined trace on X, and on the measurable tangent bundle of X, essentially because of the rectifiability.[5] Because $f = \bar{f}|X$ has a Lipschitz inverse, there exists a constant $c > 0$ such that

$$*(\rho_1 \wedge \cdots \wedge \rho_n) \geq c > 0 \tag{6.2}$$

almost everywhere on X, where the Hodge star operator $*$ is determined by the chosen orientation on X.

Condition (6.2) was turned into a definition in [17]. We say that X admits *local Cartan-Whitney presentations* if for each point $p \in X$ one can find an n-tuple of flat 1-forms $\rho = (\rho_1, \ldots, \rho_n)$ defined in an $\mathbf{R}^N$-neighborhood of p such that condition (6.2) is satisfied on X near the point p.

Theorem 1 [17] *Let $X \subset \mathbf{R}^N$ be a metrically n-dimensional, linearly locally contractible n-manifold admitting local Cartan-Whitney presentations. Then X is locally bi-Lipschitz equivalent to $\mathbf{R}^n$ outside a closed set of measure zero and of topological dimension at most $n - 2$.*

To prove Theorem 1, fix a point $p \in X$, and let $\rho = (\rho_1, \ldots, \rho_n)$ be a Cartan-Whitney presentation near p. The requirement that ρ be flat together with inequality (6.2) can be seen as a quasiregularity condition for forms.[6] We define a mapping

$$f(x) = \int_{[p,x]} \rho \tag{6.3}$$

for x sufficiently near p, where $[p, x]$ is the line segment in $\mathbf{R}^N$ from p to x, and claim that Reshetnyak's program can be run under the stipulated conditions on X. In particular, we show that for a sufficiently small neighborhood U of p in X, the map $f : U \to \mathbf{R}^n$ given in (6.3) is a branched covering which is locally bi-Lipschitz outside its branch set B_f, which furthermore is of measure zero and of topological dimension at most $n-2$. It is important to note that ρ is not assumed to be closed, so that $df \neq \rho$ in general.

[5] There is a technical point about orientation which we ignore here [17, 3.26].

[6] In fact, (6.2) resembles a stronger, Lipschitz version of (2.1) studied in [26], [40], [15].

In executing Reshetnyak's proof, we use recent advances of differential analysis on nonsmooth spaces [13], [20], [36], as well as the theory developed simultaneously in [15]. Incidentally, we avoid the use of the Harnack inequality for solutions, and therefore a deeper use of equation (2.2); this small improvement to Reshetnyak's argument was found earlier in a different context in [12].

Theorem 1 provides bi-Lipschitz coordinates for X only on a dense open set. In general, one cannot have more than that. The double suspension of a homology 3-sphere, as discussed in Section 4, can be mapped to the standard 5-sphere by a finite-to-one, piecewise linear sense-preserving map. By pulling back the standard coframe by such map, we obtain a global Cartan-Whitney presentation on a space that is not locally bi-Lipschitz equivalent to $\mathbf{R}^5$. Similar examples in dimension $n = 3$ were constructed in [14], [15], by using Semmes's spaces [34], [35]. On the other hand, we have the following result:

Theorem 2 *Let $X \subset \mathbf{R}^N$ be a metrically 2-dimensional, linearly locally contractible 2-manifold admitting local Cartan-Whitney presentations. Then X is locally bi-Lipschitz equivalent to $\mathbf{R}^2$.*

Theorem 2 is an observation of M. Bonk and myself. We use Theorem 1 together with the observation that, in dimension $n = 2$, the branch set consists of isolated points, which can be resolved. The resolution follows from the measurable Riemann mapping theorem together with the recent work by Bonk and Kleiner [4]. While Theorem 2 presents a characterization of surfaces in Euclidean space that admit local bi-Lipschitz coordinates, we do not know whether the stipulation about the existence of local Cartan-Whitney presentations is really necessary (compare Problem 1 and the discussion preceding it).

For dimensions $n \geq 3$, it would be interesting to know when there is no branching in the map (6.3). In [17], we ask if this be the case when the flat forms (ρ_i) of the Cartan-Whitney presentation belong to a Sobolev space $H^{1,2}_{loc}$ on X. The relevant example here is the map $(r, \theta, z) \mapsto (r, 2\theta, z)$, in the cylindrical coordinates of $\mathbf{R}^n$, which pulls back the standard coframe to a frame that lies in the Sobolev space $H^{1,2-\epsilon}_{loc}$ for each $\epsilon > 0$. Indeed, it was shown in [11] that in $\mathbf{R}^n$ every (Cartan-Whitney) pullback frame in $H^{1,2}_{loc}$ must come from a locally injective mapping.

7. Other recent progress and open problems

In his 1978 ICM address, Väisälä [44] asked whether the branch set of a C^1-smooth quasiregular mapping is empty if $n \geq 3$. It was known that $C^{n/(n-2)}$-smooth quasiregular mappings have no branching when $n \geq 3$. The proof in [Ri, p. 12] of this fact uses quasiregularity in a rather minimal way. In this light, the following recent result may appear surprising:

Theorem 3 [1] *For every $\epsilon > 0$ there exists a degree two $C^{3-\epsilon}$-smooth quasiregular mapping $f : \mathbf{S}^3 \to \mathbf{S}^3$ with branch set homeomorphic to $\mathbf{S}^1$.*

We are also able to improve the previous results as follows:

Theorem 4 [1] *Given $n \geq 3$ and $K \geq 1$, there exist $\epsilon = \epsilon(n, K) > 0$ and $\epsilon' = \epsilon'(n, K) > 0$ such that the branch set of every K-quasiregular mapping in a domain in $\mathbf{R}^n$ has Hausdorff dimension at most $n - \epsilon$, and that every $C^{n/(n-2)-\epsilon'}$-smooth K-quasiregular mapping in a domain in $\mathbf{R}^n$ is a local homeomorphism.*

The second assertion in Theorem 4 follows from the first, by way of Sard-type techniques. The first assertion was known earlier in a local form where $\epsilon > 0$ was dependent on the local degree [31]. Our improvement uses [31] together with the work [30] by Rickman and Srebro.

The methods in [1] fall short in showing the sharpness of Theorem 4 in dimensions $n \geq 4$ in two technical aspects. First, we would need to construct a quasiconformal homeomorphism of $\mathbf{R}^n$ to itself that is uniformly expanding on a codimension two affine subspace; moreover, such a map needs to be smooth outside this subspace. In $\mathbf{R}^3$, it is easier to construct a mapping with expanding behavior on a line; moreover, every quasiconformal homeomorphism in dimension three can be smoothened (with bounds) outside a given closed set [19].

We finish with some open problems related to branching and quasiregular mappings. The problems are neither new nor due to the author.

Problem 2 What are the possible values for the topological dimension of the branch set of a quasiregular mapping?

By suspending a covering map $H^3 \to \mathbf{S}^3$, where H^3 is as in Section 4, and using Edwards's theorem, one finds that there exists a branched covering $\mathbf{S}^5 \to \mathbf{S}^5$ that branches exactly on $\mathbf{S}^1 \subset \mathbf{S}^5$. It is not known whether there exists a quasiregular mapping $\mathbf{S}^5 \to \mathbf{S}^5$ with similar branch set. If no such map existed, we would have an interesting implication to a seemingly unrelated parametrization problem; it would follow that no double suspension of a homology 3-sphere with nontrivial fundamental group admits a *quasisymmetric* homeomorphism onto the standard 5-sphere, cf. [38], [16, Question 12].

By work of Bonk and Kleiner [4], the bi-Lipschitz parametrization problem in dimension $n = 2$ is equivalent to an analytic problem of characterizing, up to a bounded factor, the Jacobian determinants of quasiconformal mappings in $\mathbf{R}^2$. An affirmative answer to Problem 1 in Section 5 would give an affirmative answer to the following problem.

Problem 3 (Compare [16, Question 2]) Is every A_1-weight in $\mathbf{R}^2$ locally comparable to the Jacobian determinant of a quasiconformal mapping?

An *A_1-weight* is a nonnegative locally integrable function whose mean-value over each ball is comparable to its essential infimum over the ball. See [7], [32], [3], [16] for further discussion of this and related problems.

Problem 4 [16, Question 28] Is there a branched covering $f : \mathbf{S}^n \to \mathbf{S}^n$, for some $n \geq 3$, such that for every pair of homeomorphisms $\phi, \psi : \mathbf{S}^n \to \mathbf{S}^n$, the mapping $\phi \circ f \circ \psi$ fails to be quasiregular?

Branched coverings constructed by using the double suspension are obvious candidates for such mappings. In [15, 9.1], we give an example of a branched covering $f : \mathbf{S}^3 \to \mathbf{S}^3$ such that for every homeomorphism $\psi : \mathbf{S}^3 \to \mathbf{S}^3$, $f \circ \psi$ fails to be quasiregular. The example is based on a geometric decomposition space arising from *Bing's double* [34].

We close this article by commenting on the lack of direct proofs for some fundamental properties of quasiregular mappings related to branching. For example, it is known that for each $n \geq 3$ there exists $K(n) > 1$ such that every $K(n)$-quasiregular mapping is a local homeomorphism [25], [28, p. 232]. All known proofs for this fact are indirect, exploiting the Liouville theorem, and in particular there is no numerical estimate for $K(n)$. It has been conjectured that the winding mapping $(r, \theta, z) \mapsto (r, 2\theta, z)$ is the extremal here (cf. Section 6). Thus, if one uses the *inner dilatation* $K_I(f)$ of a quasiregular mapping, then conjecturally $K_I(f) < 2$ implies that $B_f = \emptyset$ for a quasiregular mapping f in $\mathbf{R}^n$ for $n \geq 3$ [29, p. 76].

Ostensibly different, but obviously a related issue, arises in search of *Bloch's constant* for quasiregular mappings. Namely, by exploiting normal families, Eremenko [10] recently proved that for given $n \geq 3$ and $K \geq 1$, there exists $b_0 = b_0(n, K) > 0$ such that every K-quasiregular mapping $f : \mathbf{R}^n \to \mathbf{S}^n$ has an inverse branch in some ball in $\mathbf{S}^n$ of radius b_0. No numerical estimate for b_0 is known. More generally, despite the deep results on value distribution of quasiregular mappings, uncovered by Rickman over the past quarter century, the affect of branching on value distribution is unknown, cf. [29, p. 96].

References

[1] M. Bonk and J. Heinonen, in preparation.

[2] M. Bonk, J. Heinonen, and S. Rohde, *Doubling conformal densities*, J. reine angew. Math., **541** (2001), 117–141.

[3] M. Bonk, J. Heinonen, and E. Saksman, *The quasiconformal Jacobian problem*, preprint (2002).

[4] M. Bonk and B. Kleiner, *Quasisymmetric parametrizations of two-dimensional metric spheres*, Inventiones Math., (to appear).

[5] M. Bonk and U. Lang, *Bi-Lipschitz parametrizations of surfaces*, in preparation.

[6] J. W. Cannon, *The characterization of topological manifolds of dimension* $n \geq 5$, Proceedings ICM (Helsinki, 1978), Acad. Sci. Fenn. Helsinki, (1980), 449–454.

[7] G. David and S. Semmes, *Strong* A_∞ *weights, Sobolev inequalities and quasiconformal mappings*, in Analysis and partial differential equations, Lecture Notes in Pure and Appl. Math., 122, Dekker, New York (1990), 101–111.

[8] S. K. Donaldson and D. P. Sullivan, *Quasiconformal* 4*-manifolds*, Acta Math., **163** (1989), 181–252.

[9] R. D. Edwards, *The topology of manifolds and cell-like maps*, Proceedings ICM (Helsinki, 1978) Acad. Sci. Fenn. Helsinki, (1980), 111–127.

[10] A. Eremenko, *Bloch radius, normal families and quasiregular mappings*, Proc. Amer. Math. Soc., **128** (2000), 557–560.

[11] J. Heinonen and T. Kilpeläinen, *BLD-mappings in $W^{2,2}$ are locally invertible*, Math. Ann., **318** (2000), 391–396.

[12] J. Heinonen and P. Koskela, *Sobolev mappings with integrable dilatation*, Arch. Rational Mech. Anal., **125** (1993), 81–97.

[13] J. Heinonen and P. Koskela, *Quasiconformal maps in metric spaces of controlled geometry*, Acta Math., **181** (1998), 1–61.

[14] J. Heinonen and S. Rickman, *Quasiregular maps $\mathbf{S}^3 \to \mathbf{S}^3$ with wild branch sets*, Topology, **37** (1998), 1–24.

[15] J. Heinonen and S. Rickman, *Geometric branched covers between generalized manifolds*, Duke Math. J., (to appear).

[16] J. Heinonen and S. Semmes, *Thirty-three yes or no questions about mappings, measures, and metrics*, Conform. Geom. Dyn., **1** (1997), 1–12.

[17] J. Heinonen and D. Sullivan, *On the locally branched Euclidean metric gauge*, Duke Math. J., (to appear).

[18] T. Iwaniec and G. Martin, *Geometric function theory and non-linear analysis*, Oxford Mathematical Monographs, Oxford University Press, Oxford (2001).

[19] M. Kiikka, *Diffeomorphic approximation of quasiconformal and quasisymmetric homeomorphisms*, Ann. Acad. Sci. Fenn. Ser. A I Math., 8 (1983), 251–256.

[20] P. Koskela, *Sobolev spaces and quasiconformal mappings on metric spaces*, Proceedings ECM (Barcelona, 2000) Progress in Math., **201**, Birkhäuser (2001), 457–467.

[21] T. J. Laakso, *Plane with A_∞-weighted metric not bilipschitz embeddable to R^n*, Bull. London Math. Soc., (to appear).

[22] J. Luukkainen and J. Väisälä, *Elements of Lipschitz topology*, Ann. Acad. Sci. Fenn. Ser. A I Math., **3** (1977), 85–122.

[23] O. Martio and V. I. Ryazanov, *The Chernavskiĭ theorem and quasiregular mappings*, Siberian Adv. Math., 10 (2000), 16–34.

[24] O. Martio and S. Rickman, *Measure properties of the branch set and its image of quasiregular mappings*, Ann. Acad. Sci. Fenn. Ser. A I Math., **541** (1973), 1–15.

[25] O. Martio, S. Rickman, and J. Väisälä, *Topological and metric properties of quasiregular mappings*, Ann. Acad. Sci. Fenn. Ser. A I Math., **488** (1971), 1–31.

[26] O. Martio and J. Väisälä, *Elliptic equations and maps of bounded length distortion*, Math. Ann., **282** (1988), 423–443.

[27] Yu. G. Reshetnyak, *Estimates of the modulus of continuity for certain mappings*, Sibirsk. Mat. Z., **7** (1966), 1106–1114; English transl. in Siberian Math. J., **7** (1966), 879–886.

[28] Yu. G. Reshetnyak, *Space mappings with bounded distortion*, Translation of Mathematical Monographs, **73**, American Mathematical Society, Providence (1989).

[29] S. Rickman, *Quasiregular mappings*, Ergebnisse der Mathematik und ihrer Grenzgebiete, **26** Springer-Verlag, Berlin Heidelberg New York (1993).

[30] S. Rickman and U. Srebro, *Remarks on the local index of quasiregular mappings*,

J. Analyse Math., 46 (1986), 246–250.
[31] J. Sarvas, *The Hausdorff dimension of the branch set of a quasiregular mapping*, Ann. Acad. Sci. Fenn. Ser. A I Math., 1 (1975), 297–307.
[32] S. Semmes, *Bi-Lipschitz mappings and strong A_∞ weights*, Ann. Acad. Sci. Fenn. Ser. A I Math., 18 (1993), 211–248.
[33] S. Semmes, *Finding structure in sets with little smoothness*, Proceedings ICM (Zürich, 1994) 1994) Birkhäuser, Basel (1995), 875–885.
[34] S. Semmes, *Good metric spaces without good parameterizations*, Rev. Mat. Iberoamericana, **12** (1996), 187–275.
[35] S. Semmes, *On the nonexistence of bi-Lipschitz parameterizations and geometric problems about A_∞ weights*, Rev. Mat. Iberoamericana, **12** (1996), 337–410.
[36] S. Semmes, *Finding curves on general spaces through quantitative topology, with applications to Sobolev and Poincaré inequalities*, Selecta Math. (N.S.), **2** (1996), 155–295.
[37] S. Semmes, *Real analysis, quantitative topology, and geometric complexity*, Publ. Mat., **45** (2001), 265–333.
[38] L. Siebenmann and D. Sullivan, *On complexes that are Lipschitz manifolds*, in Geometric topology (Proceedings Georgia Topology Conf., Athens, Ga. 1977). Edited by J. C. Cantrell, Academic Press, New York, N.Y. - London (1979), 503–525.
[39] D. Sullivan, *Hyperbolic geometry and homeomorphisms*, in Geometric topology (Proceedings Georgia Topology Conf., Athens, Ga. 1977). Edited by J. C. Cantrell, Academic Press, New York, N.Y. - London (1979), 543–555.
[40] D. Sullivan, *The exterior d, the local degree, and smoothability*, in "Prospects of Topology" (F. Quinn, ed.) Princeton Univ. Press, Princeton, New Jersey (1995).
[41] D. Sullivan, *On the foundation of geometry, analysis, and the differentiable structure for manifolds*, in "Topics in Low-Dimensional Topology" (A. Banyaga, et. al. eds.) World Scientific, Singapore-New Jersey-London-Hong Kong (1999), 89–92.
[42] T. Toro, *Surfaces with generalized second fundamental form in L^2 are Lipschitz manifolds*, J. Diff. Geom., **39** (1994), 65–101.
[43] T. Toro, *Geometric conditions and existence of bi-Lipschitz parameterizations*, Duke Math. J., **77** (1995), 193–227.
[44] J. Väisälä, *A survey of quasiregular maps in $\mathbf{R}^n$*, Proceedings ICM (Helsinki, 1978) Acad. Sci. Fenn. Helsinki, (1980), 685–691.
[45] H. Whitney *Geometric Integration Theory* Princeton University Press, Princeton, New Jersey (1957).

ICM 2002 · Vol. II · 701–709

Harmonic Measure and "Locally Flat" Domains*

Carlos E. Kenig†

Abstract

We will review work with Tatiana Toro yielding a characterization of those domains for which the harmonic measure has a density whose logarithm has vanishing mean oscillation.

2000 Mathematics Subject Classification: 31B25, 35R35, 42B35, 51M25.
Keywords and Phrases: Harmonic measure, Locally flat domains, Vanishing mean oscillation.

In this lecture, I will describe a series of joint works with Tatiana Toro on the relationship between regularity properties of harmonic measure and Poisson kernels, and regularity properties of the underlying domains. Thus, consider a domain $\Omega \subseteq \mathbb{R}^{n+1}$ and the solution to the classical Dirichlet problem:

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u\big|_{\partial\Omega} = f \in C_b(\partial\Omega), \end{cases} \tag{DP}$$

$u \in C_b(\overline{\Omega})$, where C_b is the class of bounded continuous functions. The maximum principle and the Riesz representation theorem yield the formula

$$u(X_*) = \int_{\partial\Omega} f(Q) d\omega^{X_*}(Q), \quad X_* \in \Omega,$$

and the family of positive Borel probability measures $\{d\omega^{X_*}\}$ is called harmonic measure. We sometimes fix $X_* \in \Omega$ and write $d\omega = d\omega^{X_*}$. Note that, if Ω is a smooth domain, then $d\omega^{X_*}(Q) = \frac{\partial G}{\partial \vec{n}_Q}(Q, X_*) d\sigma(Q)$, where G is the Green's function for Ω, $d\sigma$ is surface measure, and $\frac{\partial}{\partial \vec{n}_Q}$ denotes differentiation along the outward unit normal. When Ω is unbounded and v is a minimal harmonic function

*Partially supported by the NSF.
†University of Chicago, Department of Mathematics, Chicago, IL 60637, USA. E-mail: cek@math.uchicago.edu

in Ω with $v\big|_{\partial\Omega} \equiv 0$, we define $d\omega^\infty$, harmonic measure with pole at infinity, to be the measure satisfying

$$\int_{\partial\Omega} \varphi d\omega^\infty = \int_\Omega v\Delta\varphi, \quad \text{for } \varphi \in C_0^\infty(\Omega).$$

The existence and uniqueness of v and ω^∞ (modulo multiplicative constants) can be established, for instance, when Ω is an unbounded NTA (non-tangentially accessible) domain abbreviation is clarified later, it might as well be here. (see [16] for details). For example, if $\Omega = \mathbb{R}^{n+1}_+ = \{(x,t) : t > 0\}$, then $v(x,t) = t$ and $d\omega^\infty = dx$ on $\mathbb{R}^n$. The work I will describe originated from trying to understand, as $\alpha \to 0$, the classical theorem of Kellogg, which shows that, if Ω is of class $C^{1,\alpha}$, $0 < \alpha < 1$, then $d\omega = k\, d\sigma$ with $\log k \in C^\alpha$; and its "converse", the free boundary regularity of Alt-Caffarelli [1], which states that, if Ω satisfies certain necessary weak conditions (to be more fully explained later) and $d\omega = k\, d\sigma$ with $\log k \in C^\alpha$, then Ω must be of class $C^{1,\alpha}$.

To motivate our results, we recall real variable characterizations of $C^{1,\alpha}$ and C^α:

$\varphi \in C^{1,\alpha}(\mathbb{R}^n)$ $(0 < \alpha < 1) \Leftrightarrow \forall r > 0, x_0 \in \mathbb{R}^n$, there exists an affine function

$$L_{r,x_0} \text{ on } \mathbb{R}^n \text{ such that } \frac{|\varphi(x) - L_{r,x_0}(x)|}{r} \le Cr^\alpha \text{ for } |x - x_0| < r. \quad (\mathrm{I})_\alpha$$

When $\alpha = 0$, this condition is equivalent to the Zygmund class condition $\varphi \in \Lambda_*$, i.e.,

$$\frac{|\varphi(x+h) + \varphi(x-h) - 2\varphi(x)|}{|h|} \le C.$$

For us, when $\alpha = 0$, the λ_* class will also be relevant, where $\varphi \in \lambda_*$ if $\varphi \in \Lambda_*$ and, in addition, the ratio described above tends to 0 as $\alpha \to 0$.

$$h \in C^\alpha \Leftrightarrow \sup_{r>0} \frac{1}{r^\alpha} \mathrm{av}_{B_R} |h - h_{B_r}| \le C, \quad (\mathrm{II})_\alpha$$

where av_A denotes the average over the set A and B_r any ball of radius r. When $\alpha = 0$, this becomes the BMO space of John-Nirenberg [11], but we will be more interested in VMO, where $h \in \mathrm{VMO}$ if $h \in \mathrm{BMO}$ and in addition $\mathrm{av}_{B_r}|h - h_{B_r}| \xrightarrow{r\to 0} 0$. Note that VMO plays the role *vis-à-vis* BMO that continuous functions play *vis-à-vis* L^∞.

We start out by giving our geometric analogue of $(\mathrm{I})_0$: We say that $\Omega \subseteq \mathbb{R}^{n+1}$ is δ-Reifenberg flat if it has the separation property (a quantitative connectivity property) (see [16] for details), and, for all compact $K \subseteq\subseteq \mathbb{R}^{n+1}$, there exists $R_K > 0$, such that, for $0 < r < R_k$ and $Q \in \partial\Omega \cap K$, there exists an n-dimensional plane $L(r,Q)$ passing through Q such that

$$\frac{1}{r} D\big[B(r,Q) \cap \partial\Omega, B(r,Q) \cap L(r,Q)\big] \le \delta,$$

where D denotes Hausdorff distance. Note that this is a significant condition only for $\delta < 1$. We will always assume $\delta < \frac{1}{4\sqrt{2}}$. We say that Ω is Reifenberg vanishing if, as $r \to 0$, we can take $\delta \to 0$. For instance, the domain above the graph of a λ_* function is Reifenberg vanishing. In general, Reifenberg vanishing domains are not local graphs; they do not have tangent planes or a "surface measure". This class of domains was introduced by Reifenberg [20] in his study of the Plateau problem for minimal surfaces in higher dimensions.

In order to state our analogue of Kellogg's theorem in this setting, we need to introduce "multiplicative" analogues of $(\mathrm{I})_0$. A measure μ, supported on $\partial\Omega$, is doubling if, $\forall K \subseteq\subseteq \mathbb{R}^{n+1}$, there exists $R_K > 0$ such that, if $0 < r < R_K$, then

$$\mu\big(B(2r, Q) \cap \partial\Omega\big) \leq C\, \mu\big(B(r, Q) \cap \partial\Omega\big).$$

Such a μ is called asymptotically optimal doubling (see [], []) for details) if it is doubling and

$$\lim_{r\to 0} \inf_{Q\in\partial\Omega\cap K} \frac{\mu(B(\tau r, Q) \cap \partial\Omega)}{\mu(B(r, Q) \cap \partial\Omega)} = \lim_{r\to 0} \sup_{Q\in\partial\Omega\cap K} \frac{\mu(B(\tau r, Q) \cap \partial\Omega)}{\mu(B(r, Q) \cap \partial\Omega)} = \tau^n,$$

for $0 < \tau < 1$, $K \subseteq\subseteq \mathbb{R}^n$. For example, if Ω is of class $C^{1,\alpha}$ and $d\sigma$ denotes surface measure, then $\sigma\big(B(r, Q) \cap \partial\Omega\big) = \alpha_n r^n + O(r^{n+\alpha})$, $Q \in \partial\Omega$, and hence σ is asymptotically optimal doubling. If $\log k \in C^\alpha$, then the same is true for $d\omega = k\, d\sigma$. Our analog of Kellog's theorem is:

Theorem 1. ([15]) *If Ω is a Reifenberg vanishing domain, then ω (ω^∞) is asymptotically optimal doubling.*

The proof uses the fact that δ-Reifenberg flat domains are NTA domains ([9], [15]). One then uses the theory of the boundary behavior of harmonic functions on NTA domains ([9]) and comparisons to half-planes, using the Reifenberg vanishing condition and the maximum principle.

To understand a possible converse to Theorem 1, we recall a geometric measure theory (GMT) problem, first posed by Besicovitch: let μ be a positive Radon measure on $\mathbb{R}^{n+1}$ such that, for each $Q \in \Sigma$ (Σ the support of μ) and each $r > 0$, we have

$$\mu\big(B(r, Q)\big) = \alpha r^n, \quad \alpha > 0 \text{ fixed.} \tag{B}$$

Then, what can be said about μ? Clearly, if $d\mu = dx$ on $\mathbb{R}^n \subseteq \mathbb{R}^{n+1}$, then (B) holds. Nevertheless, in 1987, D. Preiss found the following interesting example: let Σ_C be the light cone $x_4^2 = x_1^2 + x_2^2 + x_3^2$, and $d\mu = d\sigma_{\Sigma_C}$ its surface measure. Then μ satisfies (B). Moreover, the general case of (B) is settled by the following remarkable theorem of Kowalski-Preiss [17].

Theorem. ([17]) *Let μ be a non-zero measure with property* (B), *and put $\Sigma = \operatorname{supp}\mu \subseteq \mathbb{R}^{n+1}$. If $n = 1, 2$, then $\Sigma = \mathbb{R}^n$. If $n \geq 3$, then either $\Sigma = \mathbb{R}^n$ or $\Sigma = \Sigma_C \otimes \mathbb{R}^{n-3}$, modulo rigid motions.*

The connection of the Preiss example to our problem comes from the fact ([16]) that, if $\Omega = \{x_4^2 < x_1^2 + x_2^2 + x_3^2\}$, $\Omega \subseteq \mathbb{R}^4$, then $d\omega^\infty = d\sigma_{\Sigma_C}$ (separation of

variables) and, by Preiss's result, ω^∞ is asymptotically optimal doubling, but, of course, Ω is not Reifenberg vanishing, since it is $\frac{1}{4\sqrt{2}}$-Reifenberg flat, and no better. Our converse to Theorem 1 is now:

Theorem 2. ([16]) *Assume that $\Omega \subseteq \mathbb{R}^{n+1}$ is an NTA, and that ω (ω^∞) is asymptotically optimal doubling. If $n = 1, 2$, then Ω is Reifenberg vanishing. If $n \geq 3$ and Ω is δ-Reifenberg flat, $\delta < \frac{1}{4\sqrt{2}}$, then Ω is Reifenberg vanishing.*

This is in fact a GMT result. It remains valid if ω (ω^∞) is replaced by any asymptotically optimal doubling measure μ with support $\partial\Omega$. The idea of the proof is to use a "blow-up" argument to reduce matters to the Kowalski-Preiss theorem. Further GMT results along these lines, also in the higher codimension case, were obtained by David-Kenig-Toro [4].

We now turn to the results motivated by $(\mathrm{II})_0$. Note that the unit normal $\vec{n}$ satisfies $|\vec{n}| = 1$, and so the BMO condition on it is automatic, but the VMO condition is not. To put our work in perspective, we recall some of the history of the subject.

A domain $\Omega \subseteq \mathbb{R}^{1+1} = \mathbb{R}^2$ is called a chord-arc domain is $\partial\Omega$ is locally rectifiable, and, whenever $Q_1, Q_2 \in \partial\Omega$, we have $\ell\big(s(Q_1, Q_2)\big) \leq C|Q_1 - Q_2|$, where ℓ denotes length and $s(Q_1, Q_2)$ is the shortest arc between Q_1 and Q_2. Ω is called vanishing chord-arc if, in addition, as $Q_1 \to Q_2$, the ratio $\frac{\ell(s(Q_1,Q_2))}{|Q_1-Q_2|}$ tends to 1, uniformly on compact sets. The first person to study harmonic measure on chord-arc domains in the plane was Lavrentiev ([18]), who proved:

Theorem. ([18]) *If $\Omega \subseteq \mathbb{R}^{1+1}$ is chord-arc, then $d\omega = k\, d\sigma$ with $\log k \in \mathrm{BMO}(d\sigma)$. (In fact, $\omega \in A_\infty(d\sigma)$, the Muckenhoupt class [7].)*

For vanishing chord-arc domains in the plane, Pommerenke [19] proved:

Theorem. ([19]) *Suppose that Ω is a chord-arc domain in $\mathbb{R}^{1+1}$. Then Ω is vanishing chord-arc if and only if $d\omega = k\, d\sigma$ with $\log k \in \mathrm{VMO}(d\sigma)$.*

These results were obtained using function theory, so their proofs don't generalize to higher dimensions. In higher dimensions, the first breakthrough came in the celebrated theorem of B. Dahlberg [2], who showed that, if $\Omega \subseteq \mathbb{R}^{n+1}$ is a Lipschitz domain, then $d\omega = k\, d\sigma$ with $\log k \in \mathrm{BMO}$ (in fact, $\omega \in A_\infty(d\sigma)$). One direction of Pommerenke's result was extended to higher dimensions by Jerison-Kenig [10], who showed that, if Ω is a C^1 domain, then $\log k \in \mathrm{VMO}$. (In general, note that Ω is of class C^1 need not imply that $\log k$ is continuous.) In order to explain our results and to clarify the connection with condition $(\mathrm{II})_0$, we need to introduce some terminology. A domain $\Omega \subseteq \mathbb{R}^{n+1}$ will be calld a chord-arc domain if it is an NTA domain (see [9]) of locally finite perimeter (see [6]) and its boundary is Ahlfors regular, i.e., the surface measure σ (which is Radon measure on $\partial\Omega$, by the assumption of locally finite perimeter) satisfies the inequalities

$$C^{-1} r^n \leq \sigma\big(B(r, Q) \cap \partial\Omega\big) \leq C r^n$$

(for $Q \in K \cap \partial\Omega$, $K \subseteq\subseteq \mathbb{R}^{n+1}$ and small r; or, if Ω is an unbounded NTA, for all $Q \in \partial\Omega$ and $r > 0$). A fundamental result of David-Jerison [3] and Semmes

[21] is that Dahlberg's theorem extends to this case, i.e., that $d\omega = k\,d\sigma$ with $\log k \in \mathrm{BMO}$, and, in fact, $\omega \in A_\infty(d\sigma)$.

We say that $\Omega \subseteq \mathbb{R}^{n+1}$ is a "δ-chord-arc domain" if Ω if δ-Reifenberg flat, Ω is of locally finite perimeter, the boundary of Ω is Ahlfors regular and the BMO norm of the unit normal $\vec{n}$ is bounded by δ. We say that Ω is "vanishing chord-arc" if, in addition, it is Reifenberg vanishing and $\vec{n} \in \mathrm{VMO}(d\sigma)$.

Remark. S. Semmes [22] has proved that, if Ω is a δ-chord-arc domain (under the definition used above), then

$$(1+\sqrt{\delta})^{-1}\alpha_n r^n \leq \sigma\big(B(r,Q)\cap\partial\Omega\big) \leq (1+\sqrt{\delta})\alpha_n r^n,$$

where α_n is the volume of the unit ball in $\mathbb{R}^n$ and $\delta \leq \delta_n$. Moreover, a combination of the results in [22] and [16] shows that, if Ω is a δ-Reifenberg flat domain which is of locally finite perimeter, and for which $\sigma\big(B(r,Q)\cap\partial\Omega\big) \leq (1+\delta)\alpha_n r^n$, then the BMO norm of $\vec{n}$ is bounded by $C\sqrt{\delta}$ for $\delta < \delta_n$. Thus, the two notions introduced of "vanishing chord-arc" domains in the plane are the same, and a domain is vanishing chord-arc exactly when it is of locally finite perimeter, has an Ahlfors regular boundary, it is Reifenberg vanishing and satisfies $\vec{n} \in \mathrm{VMO}$.

Our potential-theoretic result, which extends the work of Jerison-Kenig [10], is

Theorem 3. ([15]) *If Ω is a vanishing chord-arc domain, then ω (ω^∞) has the property that $d\omega = k\,d\sigma$ ($d\omega^\infty = h\,d\sigma$) with $\log k \in \mathrm{VMO}$ ($\log h \in \mathrm{VMO}$).*

This was proved by a combination of real variable arguments, potential-theoretic arguments, and the estimates in [10].

In order to understand possible converses of this, extending the work of Pommerenke to higher dimensions, we will recall precisely the Alt-Caffarelli [1] result which we alluded to earlier. In the language that we have introduced, their local regularity theorem can be stated as follows:

Theorem. ([1]) *Let Ω be a set of locally finite perimeter whose boundary is Ahlfors regular. Assume that Ω is δ-Reifenberg flat, $\delta < \delta_n$. Suppose that $d\omega = k\,d\sigma$ with $\log k \in C^\alpha(\partial\Omega)$ ($0 < \alpha < 1$). Then Ω is a $C^{1,\alpha}$ domain.*

The reason for this being a free boundary regularity result is that, in the case when Ω is unbounded and $d\omega = d\omega^\infty$, $d\omega^\infty = h\,d\sigma$, then $v > 0$ in Ω, $v\big|_{\partial\Omega} \equiv 0$, $\Delta v = 0$ in Ω and $h = \frac{\partial v}{\partial \vec{n}}$. Thus, knowledge of the regularity of the Cauchy data of v ($v\big|_{\partial\Omega}$, $\frac{\partial v}{\partial \vec{n}}\big|_{\partial\Omega}$) yields regularity of $\partial\Omega$ (or of $\vec{n}$, the normal).

The first connection between the above Theorem and the work of Pommerenke was made by Jerison [8], who was also the first to formulate the higher-dimensional analogues of Pommerenke's theorem as end-point estimates as $\alpha \to 0$ in the Alt-Caffarelli theorem. Jerison's theorem in [8] states that, if Ω is a Lipschitz domain and $d\omega = k\,d\sigma$ with $\log k$ continuous, then $\vec{n} \in \mathrm{VMO}$. There is an error in Lemma 4 of Jerison's paper. Nonetheless, in [16] we made strong use of the ideas in [8]. In the more recent version of our results [14], we bypass this approach.

Before stating our result, it is useful to classify the assumptions in the Alt-Caffarelli theorem. For this, we recall some examples:

Examples. When $n = 1$, Keldysh-Lavrentiev [12] (see also [5]) constructed domains in $\mathbb{R}^{1+1}$ with locally rectifiable boundaries which ([5]) can be taken to be Reifenberg vanishing and for which $d\omega = d\sigma$, i.e., $k \equiv 1$, but which are not very smooth. For instance, they fail to be chord-arc. These domains do not, of course, have Ahlfors regular boundaries. When $n = 2$, Alt-Caffarelli constructed a double cone Γ in $\mathbb{R}^3$ such that, for Ω the domain outside the cone, $d\omega^\infty = d\sigma$, i.e., $k \equiv 1$. This is of course not smooth near the origin, the problem being that, while Ω is NTA and $\partial\Omega$ is Ahlfors regular, Ω is not δ-Reifenberg flat for small δ. When $n = 3$, the Preiss cone we saw before exhibits the same behavior.

Our first result was:

Theorem 4. ([16]) *Assume that $\Omega \subseteq \mathbb{R}^{n+1}$ is δ-chord-arc, $\delta \leq \delta_n$, that ω (ω^∞) is asymptotically optimal doubling and that $\log k \in \mathrm{VMO}$ ($\log h \in \mathrm{VMO}$). Then $\vec{n} \in \mathrm{VMO}$ and Ω is vanishing chord-arc.*

Notice, however, that, when comparing the hypothesis of Theorem 4 to the Alt-Caffarelli theorem two things are apparent : first, we are making the additional assumption that ω is asymptotically optimal doubling, and hence, in light of Theorem 2, Ω is Reifenberg vanishing. Next, the "flatness" assumption in the Alt-Caffarelli theorem is δ-Reifenberg flatness, while in Theorem 4 we make the *a priori* assumption that, in addition, the BMO norm of $\vec{n}$ is smaller than δ. R...". This does not make much sense. ecently we have developed a new approach which has removed these objections. We have:

Theorem 5. ([14]) *Let Ω be a set of locally finite perimeter whose boundary is Ahlfors regular. Assume that Ω is δ-Reifenberg flat, $\delta < \delta_n$. Suppose that $d\omega = k\, d\sigma$ ($d\omega^\infty = h\, d\sigma$) with $\log k \in \mathrm{VMO}(d\sigma)$ ($\log h \in \mathrm{VMO}(d\sigma)$). Then $\vec{n} \in \mathrm{VMO}(d\sigma)$ and Ω is a vanishing chord-arc domain.*

Note that Theorems 3 and 5 together give a complete characterization of the vanishing chord-arc domains in terms of their harmonic measure, in analogy with Pommerenke's 2-dimensional result, thus answering a question posed by Semmes [21].

Our technique for the proof of Theorem 5 is to use a suitable "blow-up" to reduce matters to the following version of the "Liouville theorem" of Alt-Caffarelli ([1], [13]):

Theorem 6. ([1], [13]) *Let Ω be a set of locally finite perimeter whose boundary is (unboundedly) Ahlfors regular. Assume that Ω is an unbounded δ-Reifenberg flat domain, $\delta < \delta_n$. Suppose that u and h satisfy:*

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u > 0 & \text{in } \Omega \ u\big|_{\partial\Omega} \equiv 0 \end{cases}$$

and

$$\int_\Omega u\Delta\varphi = \int_{\partial\Omega} \varphi h\, d\sigma, \quad \text{for } \varphi \in C_0^\infty(\mathbb{R}^{n+1}).$$

Suppose that $\sup_{x\in\Omega}|\nabla u(x)|\leq 1$ *and* $h(Q)\geq 1$ *for* $(d\sigma$*-)a.e.* Q *on* $d\Omega$. *Then* Ω *is a half-space and* $u(x,x_{n+1})=x_{n+1}$.

This allows us to prove the crucial blow-up result, which we now describe. Let Ω be as in Theorem 5, and assume in addition that Ω is unbounded. Suppose $d\omega^\infty = h\,d\sigma$ with $\log h\in \mathrm{VMO}(d\sigma)$, and let u be the associated harmonic function. Let $Q_i\in\partial\Omega$ and assume that $Q_i\to Q_\infty\in\partial\Omega$ as $i\to\infty$ (without loss of generality, $Q_\infty=0$). Let $\{r_i\}_{i=1}^\infty$ be a sequence of positive numbers tending to 0, and put

$$\Omega_i=\frac{1}{r_i}(\Omega-Q_i),\qquad \partial\Omega_i=\frac{1}{r_i}(\partial\Omega-Q_i),$$
$$u_i(X)=\frac{1}{r_i\mathrm{av}_{B(r_i,Q_i)}h\,d\sigma}u(r_iX+Q_i)\quad\text{and}\quad d\omega_i^\infty=h_i(Q)d\sigma_i(Q),$$

where $h_i(Q)=\frac{1}{\mathrm{av}_{B(r_i,Q_i)}h\,d\sigma}h(r_iQ+Q_i)$. Then:

Theorem 7. *There exists a subsequence of* $\{\Omega_i\}$ *(which we will call again* $\{\Omega_i\}$*) satisfying:*

$$\Omega_i\to\Omega_\infty \text{ in the Hausdorff distance sense, uniformly on compact sets;} \tag{7.1}$$
$$\partial\Omega_i\to\partial\Omega_\infty \text{ in the Hausdorff distance sense, uniformly on compact sets;} \tag{7.2}$$
$$u_i\to u_\infty \text{ uniformly on compact sets} \tag{7.3}$$

and

$$\begin{cases}\Delta u_\infty=0 & \text{in } \Omega_\infty\\ u_\infty=0 & \text{in } \partial\Omega_\infty\\ u_\infty>0 & \text{in } \Omega_\infty.\end{cases} \tag{7.4}$$

Furthermore

$$\omega_i\rightharpoonup\omega_\infty \tag{7.5}$$

and

$$\sigma_i\rightharpoonup\sigma_\infty, \tag{7.6}$$

weakly as Radon measures. Here, $\sigma_\infty=\mathcal{H}^n\lfloor\partial\Omega_\infty$ *and* ω_∞ *denotes the harmonic measure of* Ω_∞ *with pole at* ∞ *(corresponding to* u_∞*). Moreover,*

$$\sup_{Z\in\Omega_\infty}|\nabla u_\infty(Z)|\leq 1 \tag{7.7}$$

and

$$h_\infty(Q)=\frac{d\omega_\infty}{d\sigma_\infty}(Q)\geq 1\quad \text{for } \mathcal{H}^n\text{-a.e. } Q\in\partial\Omega_\infty. \tag{7.8}$$

Since $\log h \in \mathrm{VMO}(\partial\Omega)$, the average $\mathrm{av}_{B(r,Q)} h\, d\sigma$ is close to the value of $\log h$ in a proportionally large subset of $B(r,Q) \cap \partial\Omega$. This remark allows us to conclude that (7.6) holds, which is crucial to the application and which fails in general under just (7.1) and (7.2).

As an immediate application of Theorems 6 and 7, we obtain that Ω_∞ is a half-plane. This already establishes that Ω is Reifenberg vanishing in Theorem 5. To establish that $\vec{n}$ is in VMO, we assume otherwise, and obtain $Q_i \to Q_\infty$, $r_i \to 0$, such that $\mathrm{av}_{B(r_i,Q_i)} |\vec{n} - \vec{n}_{B(r_i,Q_i)}|^2 d\sigma \geq \ell^2$, $\ell > 0$. We consider the corresponding blow-up sequence, and let $\vec{e}_{n+1}$ be the direction perpendicular to $\partial\Omega_\infty$. By the divergence theorem and (7.1) and (7.2), we have for $\varphi \in C_0^\infty(\mathbb{R}^{n+1})$ that

$$\lim_{i\to\infty} \int_{\partial\Omega_i} \varphi \langle \vec{n}_i, \vec{e}_{n+1} \rangle d\sigma_i = \int_{\mathbb{R}^n \times \{0\}} \varphi\, dx$$

and hence

$$\lim_{i\to\infty} \left\{ \int_{\partial\Omega_i} \varphi\, d\sigma_i - \frac{1}{2} \int_{\partial\Omega_i} \varphi |\vec{n}_i - \vec{e}_{n+1}|^2 d\sigma_i \right\} = \int_{\mathbb{R}^n \times \{0\}} \varphi\, dx,$$

so that (7.6) yields

$$\lim_{i\to\infty} \int_{\partial\Omega_i} \varphi |\vec{n}_i - \vec{e}_{n+1}|^2 d\sigma_i = 0.$$

Taking $\varphi \geq \chi_{B(1,0)}$ yields the corresponding bound for the integral on $\partial\Omega_i \cap B(1,0)$. But

$$\mathrm{av}_{B(1,0)\cap\partial\Omega_i} |\vec{n}_i - \vec{e}_{n+1}|^2 d\sigma_i = \mathrm{av}_{B(r_i,Q_i)} |\vec{n} - \vec{e}_{n+1}|^2 d\sigma,$$

and hence

$$\ell \leq \overline{\lim_{i\to\infty}} \left(\mathrm{av}_{B(r_i,Q_i)} |\vec{n} - \vec{n}_{B(r_i,Q_i)}|^2 d\sigma \right)^{1/2} \leq 2\, \overline{\lim_{i\to\infty}} \left(\mathrm{av}_{B(r_i,Q_i)} |\vec{n} - \vec{e}_{n+1}|^2 d\sigma \right)^{1/2},$$

a contradiction. This concludes the proof.

References

[1] H. W. Alt and L. A. Caffarelli, Existence and regularity for a minimum problem with free boundary, *J. Reine Angew. Math.*, 325 (1981), 105–144.

[2] B. Dahlberg, On estimates for harmonic measure, *Arch. Rat. Mech. Anal.*, 65 (1977), 272–288.

[3] G. David and D. Jerison, Lipschitz approximation to hypersurfaces, harmonic measure and singular integrals, *Indiana Univ. Math. J.*, 39 (1990), 831–845.

[4] G. David, C. Kenig, and T. Toro, Asymptotically optimally doubling measures and Reifenberg flat sets with vanishing constant, *CPAM*, 54 (2001), 385–449.

[5] P. Duren, *The theory of H^p spaces*, Academic Press, New York, 1970.

[6] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, Studies in Advanced Mathematics, CRC Press, 1992.

[7] J. García-Cuerva and J. L. Rubio de Francia, *Weighted norm inequalities and related topics*, Math. Studies, no. 116, North Holland, 1985.

[8] D. Jerison, Regularity of the Poisson kernel and free boundary problems, *Colloquium Mathematicum*, 60–61 (1990), 547–567.
[9] D. Jerison and C. Kenig, Boundary behavior of harmonic functions in nontangentially accessible domains, *Adv. in Math.*, 46 (1982), 80–147.
[10] ———, The logarithm of the Poisson kernel of a C^1 domain has vanishing mean oscillation, *Trans. Amer. Math. Soc.*, 273 (1982), 781–794.
[11] F. John and L. Nirenberg, On functions of bounded mean oscillation, *Comm. Pure Appl. Math.*, 14 (1961), 415–426.
[12] M. V. Keldysh and M. A. Lavrentiev, Sur la représentation conforme des domaines limités par des courbes rectifiables, *Ann. Sci. Ecole Norm. Sup.*, 54 (1937), 1–38.
[13] C. Kenig and T. Toro, On the free boundary regularity theorem of Alt and Caffarelli, preprint.
[14] ———, Poisson kernel characterization of Reifenberg flat chord-arc domains, to appear, *Ann. Sci. Ec. Norm. Sup.*
[15] ———, Harmonic measure on locally flat domains, *Duke Math. J.*, 87 (1997), 509–551.
[16] ———, Free boundary regularity for harmonic measures and Poisson kernels, *Ann. of Math.*, 150 (1999), 369–454.
[17] O. Kowalski and D. Preiss, Besicovitch-type properties of measures and submanifolds, *J. Reine Angew. Math.*, 379 (1987), 115–151.
[18] M. Lavrentiev, Boundary problems in the theory of univalent functions, *Math Sb. (N. S.) I*, 43 (1936), 815–844.
[19] Ch. Pommerenke, On univalent functions, Bloch functions and VMOA, *Math. Ann.*, 236 (1978), 199–208.
[20] E. Reifenberg, Solution of the Plateau problem for m-dimensional surfaces of varying topological type, *Acta Math.*, 104 (1960), 1–92.
[21] S. Semmes, Analysis vs. geometry on a class of rectifiable hypersurfaces, *Indiana Univ. J.*, 39 (1990), 1005–1035.
[22] ———, Chord-arc surfaces with small constant I, *Adv. in Math.*, 85 (1991), 198–293.

ICM 2002 · Vol. II · 711–720

Solving Pseudo-Differential Equations

Nicolas Lerner*

Abstract

In 1957, Hans Lewy constructed a counterexample showing that very simple and natural differential equations can fail to have local solutions. A geometric interpretation and a generalization of this counterexample were given in 1960 by L.Hörmander. In the early seventies, L.Nirenberg and F.Treves proposed a geometric condition on the principal symbol, the so-called condition (ψ), and provided strong arguments suggesting that it should be equivalent to local solvability. The necessity of condition (ψ) for solvability of pseudo-differential equations was proved by L.Hörmander in 1981. The sufficiency of condition (ψ) for solvability of differential equations was proved by R.Beals and C.Fefferman in 1973. For differential equations in any dimension and for pseudo-differential equations in two dimensions, it was shown more precisely that (ψ) implies solvability with a loss of one derivative with respect to the elliptic case: for instance, for a complex vector field X satisfying (ψ), $f \in L^2_{\rm loc}$, the equation $Xu = f$ has a solution $u \in L^2_{\rm loc}$.

In 1994, it was proved by N.L. that condition (ψ) does not imply solvability with loss of one derivative for pseudo-differential equations, contradicting repeated claims by several authors. However in 1996, N.Dencker proved that these counterexamples were indeed solvable, but with a loss of two derivatives. We shall explore the structure of this phenomenon from both sides: on the one hand, there are first-order pseudo-differential equations satisfying condition (ψ) such that no $L^2_{\rm loc}$ solution can be found with some source in $L^2_{\rm loc}$. On the other hand, we shall see that, for these examples, there exists a solution in the Sobolev space $H^{-1}_{\rm loc}$.

The sufficiency of condition (ψ) for solvability of pseudo-differential equations in three or more dimensions is still an open problem. In 2001, N.Dencker announced that he has proved that condition (ψ) implies solvability (with a loss of two derivatives), settling the Nirenberg-Treves conjecture. Although his paper contains several bright and new ideas, it is the opinion of the author of these lines that a number of points in his article need clarification.

2000 Mathematics Subject Classification: 35S05, 35A05, 47G30.
Keywords and Phrases: Solvability, Pseudo-Differential equation, Condition (ψ).

*University of Rennes, Université de Rennes 1, Irmar, Campus de Beaulieu, 35042 Rennes cedex, France. E-mail: lerner@univ-rennes1.fr

1. From Hans Lewy to Nirenberg-Treves' condition (ψ)

Year 1957.

The Hans Lewy operator L_0, introduced in [20], is the following complex vector field in $\mathbb{R}^3$

$$L_0 = \frac{\partial}{\partial x_1} + i\frac{\partial}{\partial x_2} + i(x_1 + ix_2)\frac{\partial}{\partial x_3}. \tag{1.1}$$

There exists $f \in C^\infty$ such that the equation

$$L_0 u = f \tag{1.2}$$

has no distribution solution, even locally. This discovery came as a great shock for several reasons. First of all, L_0 has a very simple expression and is natural as the Cauchy-Riemann operator on the boundary of the pseudo-convex domain

$$\{(z_1, z_2) \in \mathbb{C}^2, |z_1|^2 + 2\,\mathrm{Im}\, z_2 < 0\}.$$

Moreover L_0 is a non-vanishing vector field so that no pathological behaviour related to multiple characteristics is to be expected. In the fifties, it was certainly the conventional wisdom that any "reasonable" operator should be locally solvable, and obviously (1.1) was indeed very reasonable so the conclusion was that, once more, the CW should be revisited. One of the questions posed by such a counterexample was to find some geometric explanation for this phenomenon.

1960.

This was done in 1960 by L.Hörmander in [7] who proved that if p is the symbol of a differential operator such that, at some point (x, ξ) in the cotangent bundle,

$$p(x,\xi) = 0 \quad \text{and} \quad \{\mathrm{Re}\, p, \mathrm{Im}\, p\}(x,\xi) > 0, \tag{1.3}$$

then the operator P with principal symbol p is not locally solvable at x; in fact, there exists $f \in C^\infty$ such that, for any neighborhood V of x the equation $Pu = f$ has no solution $u \in \mathcal{D}'(V)$. Of course, in the case of differential operators, the sign > 0 in (1.3) can be replaced by $\neq 0$ since the Poisson bracket $\{\mathrm{Re}\, p, \mathrm{Im}\, p\}$ is then an homogeneous polynomial with odd degree in the variable ξ. Nevertheless, it appeared later (in [8]) that the same statement is true for pseudo-differential operators, so we keep it that way. Since the symbol of $-iL_0$ is $\xi_1 - x_2\xi_3 + i(\xi_2 + x_1\xi_3)$, and the Poisson bracket $\{\xi_1 - x_2\xi_3, \xi_2 + x_1\xi_3\} = 2\xi_3$, the assumption (1.3) is fulfilled for L_0 at any point x in the base and the nonsolvability property follows. This gives a necessary condition for local solvability of pseudo-differential equations: a locally solvable operator P with principal symbol p should satisfy

$$\{\mathrm{Re}\, p, \mathrm{Im}\, p\}(x,\xi) \leq 0 \quad \text{at} \quad p(x,\xi) = 0. \tag{1.4}$$

Naturally, condition (1.4) is far from being sufficient for solvability (see e.g. the nonsolvable M_3 below in (1.5)). After the papers [20], [7], the curiosity of the

mathematical community was aroused in search of a geometric condition on the principal symbol, characterizing local solvability of principal type operators. It is important to note that for principal type operators with a real principal symbol, such as a non-vanishing real vector field, or the wave equation, local solvability was known after the 1955 paper of L.Hörmander in [6]. In fact these results extend quite easily to the pseudo-differential real principal type case. As shown by the Hans Lewy counterexample and the necessary condition (1.4), the matters are quite different for complex-valued symbols.

1963.

It is certainly helpful to look now at some simple models. For $t, x \in \mathbb{R}$, with the usual notations

$$D_t = -i\partial_t, \quad (\widehat{|D_x|u})(\xi) = |\xi|\hat{u}(\xi),$$

where $\hat{u}$ is the x-Fourier transform of u, $l \in \mathbb{N}$, let us consider the operators defined by

$$M_l = D_t + it^l D_x, \quad N_l = D_t + it^l |D_x|. \tag{1.5}$$

It is indeed rather easy to prove that, for $k \in \mathbb{N}$, $M_{2k}, N_{2k}, N_{2k+1}^*$ are solvable whereas M_{2k+1}, N_{2k+1} are nonsolvable. In particular, the operators M_1, N_1 satisfy (1.3). On the other hand, the operator $N_1^* = D_t - it|D_x|$ is indeed solvable since its adjoint operator N_1 verifies the a priori estimate

$$T\|N_1 u\|_{L^2(\mathbb{R}^2)} \geq \|u\|_{L^2(\mathbb{R}^2)},$$

for a smooth compactly supported u vanishing for $|t| \geq T/2$. No such estimate is satisfied by $N_1^* u$ since its x-Fourier transform is

$$-i\partial_t v - it|\xi|v = (-i)(\partial_t v + t|\xi|v),$$

where v is the x-Fourier transform of u. A solution of $N_1^* u = 0$ is thus given by the inverse Fourier transform of $e^{-t^2|\xi|/2}$, ruining solvability for the operator N_1. A complete study of solvability properties of the models M_l was done in [23] by L.Nirenberg and F.Treves, who also provided a sufficient condition of solvability for vector fields; the analytic-hypoellipticity properties of these operators were also studied in a paper by S.Mizohata [21].

1971.

The ODE-like examples (1.5) led L.Nirenberg and F.Treves in [24–25–26] to formulate a conjecture and to prove it in a number of cases, providing strong grounds in its favour. To explain this, let us look simply at the operator

$$L = D_t + iq(t, x, D_x), \tag{1.6}$$

where q is a real-valued first-order symbol. The symbol of L is thus $\tau + iq(t, x, \xi)$. The bicharacteristic curves of the real part are oriented straight lines with direction $\partial/\partial t$; now we examine the variations of the imaginary part $q(t, x, \xi)$ along these lines. It amounts only to check the functions $t \mapsto q(t, x, \xi)$ for fixed (x, ξ). The

good cases in (1.5) (when solvability holds) are $t^{2k}\xi, -t^{2k+1}|\xi|$: when t increases these functions do not change sign from $-$ to $+$. The bad cases are $t^{2k+1}|\xi|$: when t increases these functions do change sign from $-$ to $+$; in particular, the nonsolvable case (1.3), tackled in [8], corresponds to a change of sign of $\operatorname{Im} p$ from $-$ to $+$ at a simple zero. The general formulation of condition (ψ) for a principal type operator with principal symbol p is as follows: for all $z \in \mathbb{C}$, $\operatorname{Im}(zp)$ does not change sign from $-$ to $+$ along the oriented bicharacteristic curves of $\operatorname{Re}(zp)$. It is a remarkable and non-trivial fact that this condition is invariant by multiplication by an elliptic factor as well as by composition with an homogeneous canonical transformation. The *Nirenberg-Treves conjecture*, proved in several cases in [24–25–26], such as for differential operators with analytic coefficients, states that, *for a principal type pseudo-differential equation, condition (ψ) is equivalent to local solvability.*

The paper [25] introduced a radically new method of proof of energy estimates for the adjoint operator L^* based on a factorization of q in (1.6): whenever

$$q(t,x,\xi) = a(t,x,\xi)b(x,\xi) \tag{1.7}$$

with $a \leq 0$ of order 0 and b of order 1, then the operator L in (1.6) is locally solvable. Looking simply at the ODE

$$D_t + ia(t,x,\xi)b(x,\xi) = (-i)\big(\partial_t - a(t,x,\xi)b(x,\xi)\big), \tag{1.8}$$

it is clear that in the region $\{b(x,\xi) \geq 0\}$, the forward Cauchy problem for (1.8) is well posed, whereas in $\{b(x,\xi) \leq 0\}$, well-posedness holds for the backward Cauchy problem. This remark led L.Nirenberg and F.Treves to use as a multiplier in the energy method the sign of the operator with symbol b. They were also able to provide the proper commutator estimates to handle the remainder terms generated by this operator-theoretic method. Although a factorization (1.7) can be obtained for differential operators with analytic regularity satisfying condition (ψ), such a factorization is not true in the C^∞ case. Incidentally, one should note that for differential operators, condition (ψ) is equivalent to ruling out any change of sign of $\operatorname{Im} p$ along the bicharacteristics of $\operatorname{Re} p$ (the latter condition is called condition (P)); this fact is due to the identity $p(x,-\xi) = (-1)^m p(x,\xi)$, valid for an homogeneous polynomial of degree m in the variable ξ.

Using the Malgrange-Weierstrass theorem on normal forms of complex-valued non-degenerate C^∞ functions and the Egorov theorem on quantization of homogeneous canonical transformations, there is no loss of generality considering only first order operators of type (1.6). The expression of condition (ψ) for L is then very simple since it reads

$$q(t,x,\xi) < 0 \quad \text{and} \quad s > t \Longrightarrow q(s,x,\xi) \leq 0. \tag{1.9}$$

Note that the expression of condition (P) for L is simply $q(t,x,\xi)q(s,x,\xi) \geq 0$. Much later in 1988, N.Lerner [14] proved the sufficiency of condition (ψ) for local solvability of pseudo-differential equations in two dimensions and as well for the classical oblique-derivative problem [15]. The method of proof of these results

is based upon a factorization analogous to (1.7) but where $b(x,\xi)$ is replaced by $\beta(t,x)|\xi|$ and β is a smooth function such that $t \mapsto \beta(t,x)$ does not change sign from $+$ to $-$ when t increases. Then a properly defined sign of $\beta(t,x)$ appears as a non-decreasing operator and the Nirenberg-Treves energy method can be adapted to this situation.

1973.

At this date, R.Beals and C.Fefferman [1] took as a starting point the previous results of L.Nirenberg and F.Treves and, removing the analyticity assumption, they were able to prove the sufficiency of condition (P) for local solvability, obtaining thus the sufficiency of condition (ψ) for local solvability of differential equations. The key ingredient was a drastically new vision of the pseudo-differential calculus, defined to obtain the factorization (1.7) in regions of the phase space much smaller than cones or semi-classical "boxes" $\{(x,\xi), |x| \leq 1, |\xi| \leq h^{-1}\}$. Considering the family $\{q(t,x,\xi)\}_{t\in[-1,1]}$ of classical homogeneous symbols of order 1, they define, via a Calderón-Zygmund decomposition, a pseudo-differential calculus depending on the family $\{q(t,\cdot)\}$, in which all these symbols are first order but also such that, at some level t_0, some ellipticity property of $q(t_0,\cdot)$ or $\nabla_{x,\xi}q(t_0,\cdot)$ is satisfied. Condition (P) then implies easily a factorization of type (1.7) and the Nirenberg-Treves energy method can be used. It is interesting to notice that some versions of these new pseudo-differential calculi were used later on for the proof of the Fefferman-Phong inequality [5]. In fact, the proof of R.Beals and C.Fefferman marked the day when microlocal analysis stopped being only homogeneous or semi-classical, thanks to methods of harmonic analysis such as Calderón-Zygmund decomposition made compatible with the Heisenberg uncertainty principle.

1978.

Going back to solvability problems, the existence of C^∞ solutions for C^∞ sources was proved by L.Hörmander in [9] for pseudo-differential equations satisfying condition (P). For such an operator P of order m, satisfying also a non-trapping condition, a semi-global existence theorem was proved, with a loss of $1+\epsilon$ derivatives, with $\epsilon > 0$. Following an idea given by R.D.Moyer [22] for a result in two dimensions, L.Hörmander proved in [10] that condition (ψ) is necessary for local solvability: assuming that condition (ψ) is not satisfied for a principal type operator P, he was able to construct approximate non-trivial solutions u for the adjoint equation $P^*u = 0$, which implies that P is not solvable. Although the construction is elementary for the model operators N_{2k+1} in (1.5) (as sketched above for N_1 in our 1963 section), the multidimensional proof is rather involved and based upon a geometrical optics method adapted to the complex case. The details can be found in the proof of theorem 26.4.7′ of [11].

We refer the reader to the paper [13] for a more detailed historical overview of this problem. On the other hand, it is clear that our interest is focused on solvability in the C^∞ category. Let us nevertheless recall that the sufficiency of condition (ψ) in the analytic category (for microdifferential operators acting on microfunctions) was proved by J.-M.Trépreau [27] (see also [12], chapter VII).

2. Counting the loss of derivatives

Condition (ψ) does not imply solvability with loss of one derivative.

Let us consider a principal-type pseudo-differential operator L of order m. We shall say that L is locally solvable with a loss of μ derivatives whenever the equation $Lu = f$ has a local solution u in the Sobolev space $H^{s+m-\mu}$ for a source f in H^s. Note that the loss is zero if and only if L is elliptic. Since for the simplest principal type equation $\partial/\partial x_1$, the loss of derivatives is 1, we shall consider that 1 is the "ordinary" loss of derivatives. When L satisfies condition (P) (e.g. if L is a differential operator satisfying condition (ψ)), or when L satisfies condition (ψ) in two dimensions, the estimates

$$C\|L^*u\|_{H^s} \geq \|u\|_{H^{s+m-1}}, \tag{2.1}$$

valid for smooth compactly supported u with a small enough support, imply local solvability with loss of 1 derivative, the ordinary loss referred to above. For many years, repeated claims were made that condition (ψ) for L implies (2.1), that is solvability with loss of 1 derivative. It turned out that these claims were wrong, as shown in [16] by the following result (see also section 6 in the survey [13]).

Theorem 2.1. *There exists a principal type first-order pseudo-differential operator L in three dimensions, satisfying condition (ψ), a sequence u_k of C_c^∞ functions with* $\operatorname{supp} u_k \subset \{x \in \mathbb{R}^3, |x| \leq 1/k\}$ *such that*

$$\|u_k\|_{L^2(\mathbb{R}^3)} = 1, \quad \lim_{k\to+\infty} \|L^*u_k\|_{L^2(\mathbb{R}^3)} = 0. \tag{2.2}$$

As a consequence, for this L, there exists $f \in L^2$ such that the equation $Lu = f$ has no local solution u in L^2. We shall now briefly examine some of the main features of this counterexample, leaving aside the technicalities which can be found in the papers quoted above. Let us try, with $(t, x, y) \in \mathbb{R}^3$,

$$L = D_t - ia(t)\big(D_x + H(t)V(x)|D_y|\big), \tag{2.3}$$

with $H = \mathbf{1}_{\mathbb{R}_+}$, $C^\infty(\mathbb{R}) \ni V \geq 0$, $C^\infty(\mathbb{R}) \ni a \geq 0$ flat at 0. Since the function $q(t, x, y, \xi, \eta) = -a(t)\big(\xi + H(t)V(x)|\eta|\big)$ satisfies (1.9) as the product of the non-positive function $-a(t)$ by the non-decreasing function $t \mapsto \xi + H(t)V(x)|\eta|$, the operator L satisfies condition (ψ). To simplify the exposition, let us assume that $a \equiv 1$, which introduces a rather unimportant singularity in the t-variable, let us replace $|D_y|$ by a positive (large) parameter Λ, which allows us to work now only with the two real variables t, x and let us set $W = \Lambda V$. We are looking for a non-trivial solution $u(t, x)$ of $L^*u = 0$, which means then

$$\partial_t u = \begin{cases} D_x u, & \text{for } t < 0, \\ \big(D_x + W(x)\big)u, & \text{for } t > 0. \end{cases}$$

The operator $D_x + W$ is unitarily equivalent to D_x: with $A'(x) = W(x)$, we have $D_x + W(x) = e^{-iA(x)} D_x e^{iA(x)}$, so that the negative eigenspace of the operator

$D_x + W(x)$ is $\{v \in L^2(\mathbb{R}), \operatorname{supp} \widehat{e^{iA}v} \subset \mathbb{R}_-\}$. Since we want u to decay when $t \to \pm\infty$, we need to choose $v_1, v_2 \in L^2(\mathbb{R})$, such that

$$u(t,x) = \begin{cases} e^{tD_x}v_1, & \operatorname{supp} \widehat{v_1} \subset \mathbb{R}_+ \quad \text{for } t < 0, \\ e^{t(D_x+W)}v_2, & \operatorname{supp} \widehat{e^{iA}v_2} \subset \mathbb{R}_- \quad \text{for } t > 0. \end{cases} \tag{2.4}$$

We shall not be able to choose $v_1 = v_2$ in (2.4), so we could only hope for L^*u to be small if $\|v_2 - v_1\|_{L^2(\mathbb{R})}$ is small. Thus this counterexample is likely to work if the unit spheres of the vector spaces

$$E_1^+ = \{v \in L^2(\mathbb{R}), \operatorname{supp} \widehat{v} \subset \mathbb{R}_+\} \quad \text{and} \quad E_2^- = \{v \in L^2(\mathbb{R}), \operatorname{supp} \widehat{e^{iA}v} \subset \mathbb{R}_-\}$$

are close. Note that since $W \geq 0$, we get $E_1^+ \cap E_2^- = \{0\}$: in fact, with $L^2(\mathbb{R})$ scalar products, we have

$$v \in E_1^+ \cap E_2^- \Longrightarrow 0 \overset{v\in E_1^+}{\leq} \langle Dv, v\rangle \overset{0\leq W}{\leq} \langle (D+W)v, v\rangle \overset{v\in E_2^-}{\leq} 0 \Longrightarrow \langle Dv, v\rangle = 0$$

which gives $v = 0$ since $v \in E_1^+$. Nevertheless, the "angle" between E_1^+ and E_2^- could be small for a careful choice of a positive W. It turns out that $W_0(x) = \pi\delta_0(x)$ is such a choice. Of course, several problems remain such as regularize W_0 in such a way that it becomes a first-order semi-classical symbol, redo the same construction with a smooth function a flat at 0 and various other things.

Anyhow, these difficulties eventually turn out to be only technical, and *in fine*, the actual reason for which theorem 2.1 is true is simply that the positive eigenspace of D_x (i.e. $L^2(\mathbb{R})$ functions whose Fourier transform is supported in $\mathbb{R}_+$) could be arbitrarily close to the negative eigenspace of $D_x + W(x)$ for some non-negative W, triggering nonsolvability in L^2 for the three-dimensional model operator

$$D_t - ia(t)\big(D_x + \mathbf{1}_{\mathbb{R}_+}(t)W(x)|D_y|\big), \tag{2.5}$$

where a is some non-negative function, flat at 0. This phenomenon is called the "drift" in [16] and could not occur for differential operators or for pseudo-differential operators in two dimensions. A more geometric point of view is that for a principal type symbol p, satisfying condition (ψ), one may have bicharacteristics of $\operatorname{Re} p$ which stay in the set $\{\operatorname{Im} p = 0\}$. This can even occur for operators satisfying condition (P). However condition (P) ensures that the nearby bicharacteristics of $\operatorname{Re} p$ stay either in $\{\operatorname{Im} p \geq 0\}$ or in $\{\operatorname{Im} p \leq 0\}$. This is no longer the case when condition (ψ) holds, although the bicharacteristics are not allowed to pass from $\{\operatorname{Im} p < 0\}$ to $\{\operatorname{Im} p > 0\}$. The situation of having a bicharacteristic of $\operatorname{Re} p$ staying in $\{\operatorname{Im} p = 0\}$ will generically trigger the drift phenomenon mentioned above when condition (P) does not hold. So the counterexamples to solvability with loss of one derivative are in fact very close to operators satisfying condition (P).

A related remark is that the ODE-like solvable models in (1.5) do not catch the generality allowed by condition (ψ). Even for subelliptic operators, whose tranposed are of course locally solvable, it is known that other model operators than M_{2k}, N_l can occur. In particular the three-dimensional models $D_t + it^{2k}(D_x + t^{2l+1}x^{2m}|D_y|)$, where k, l, m are non-negative integers are indeed subelliptic and are not reducible to (1.5) (see chapter 27 in [11] and the remark before corollary 27.2.4 there).

Solvability with loss of two derivatives.

Although theorem 2.1 demonstrates that condition (ψ) does not imply solvability with loss of one derivative, the counterexamples constructed in this theorem are indeed solvable, but with a loss of two derivatives, as proven by N.Dencker in 1996 [2]. The same author gave a generalization of his results in [3] and later on, analogous results were given in [17].

A measurable function $p(t,x,\xi)$ defined on $\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n$ will be called in the next theorem a symbol of order m whenever, for all $(\alpha,\beta)\in\mathbb{N}^n\times\mathbb{N}^n$

$$\sup_{(t,x,\xi)\in\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n}|(\partial_x^\alpha\partial_\xi^\beta p)(t,x,\xi)|(1+|\xi|)^{-m+|\beta|}<+\infty. \tag{2.6}$$

Theorem 2.2. *Let $a(t,x,\xi)$ be a non-positive symbol of order 0, $b(t,x,\xi)$ be a real-valued symbol of order 1 such that $\partial_t b\geq 0$, and $r(t,x,\xi)$ be a (complex-valued) symbol of order 0. Then the operator*

$$L=D_t+ia(t,x,D_x)b(t,x,D_x)+r(t,x,D_x) \tag{2.7}$$

is locally solvable with a loss of two derivatives. Since the counterexamples constructed in theorem 2.1 are in fact of type (2.7), they are locally solvable with a loss of two derivatives.

In fact, for all points in $\mathbb{R}^{n+1}$, there exists a neighborhood V, a positive constant C such that, for all $u\in C_c^\infty(V)$

$$C\|L^*u\|_{H^0}\geq\|u\|_{H^{-1}}. \tag{2.8}$$

This estimate actually represents a loss of two derivatives for the first-order L; the estimate with loss of 0 derivative would be $\|L^*u\|_{H^0}\gtrsim\|u\|_{H^1}$, the estimate with loss of one derivative would be $\|L^*u\|_{H^0}\gtrsim\|u\|_{H^0}$, and both are false, the first because L^* is not elliptic, the second from theorem 2.1. The proof of theorem 2.2 is essentially based upon the energy method which boils down to compute for all $T\in\mathbb{R}$

$$\mathrm{Re}\langle L^*u,iBu+iH(t-T)u\rangle_{L^2(\mathbb{R}^{n+1})}$$

where $B=b(t,x,D_x)$. Some complications occur in the proof from the rather weak assumption $\partial_t b\geq 0$ and also from the lower order terms. Anyhow, the correct multiplier is essentially given by $b(t,x,D_x)$. Theorem 2.2 can be proved for much more general classes of pseudo-differential operators than those given by (2.6). As a consequence, it can be extended naturally to contain the solvability result under condition (P) (but with a loss of two derivatives, see e.g. theorem 3.4 in [17]).

Miscellaneous results.

Let us mention that the operator (1.6) is solvable with a loss of one derivative (the ordinary loss) if condition (ψ) is satisfied (i.e. (1.9)) as well as the extra condition

$$|\partial_x q(t,x,\xi)|^2|\xi|^{-1}+|\partial_\xi q(t,x,\xi)|^2|\xi|\leq C|\partial_t q(t,x,\xi)|\quad\text{when}\quad q(t,x,\xi)=0.$$

This result is proved in [18] and shows that "transversal" changes of sign do not generate difficulties. Solvability with loss of one derivative is also true for operators satisfying condition (ψ) such that the changes of sign take place on a Lagrangean manifold, e.g. operators (1.6) such that the sign of $q(t,x,\xi)$ does not depend on ξ, i.e. $q(t,x,\xi)q(t,x,\eta) \geq 0$ for all (t,x,ξ,η). This result is proved in section 8 of [13] which provides a generalization of [15] where the standard oblique-derivative problem was tackled. On the other hand, it was proved in [19] that for a first-order pseudo-differential operator L satisfying condition (ψ), there exists a L^2 bounded perturbation R such that $L+R$ is locally solvable with loss of two derivatives.

3. Conclusion and perspectives

The following facts are known for principal type pseudo-differential operators.

F1. Local solvability implies (ψ).
F2. For differential operators and in two dimensions, (ψ) implies local solvability.
F3. (ψ) does not imply local solvability with loss of one derivative.
F4. The known counterexamples in (F3) are solvable with loss of two derivatives.

The following questions are open.

Q1. Is (ψ) sufficient for local solvability in three or more dimensions?
Q2. If the answer to Q1 is yes, what is the loss of derivatives?
Q3. In addition to (ψ), which condition should be required to get local solvability with loss of one derivative?
Q4. Is analyticity of the principal symbol and condition (ψ) sufficient for local solvability?

The most important question is with no doubt Q1, since, with F1, it would settle the Nirenberg-Treves conjecture. From F3, it appears that the possible loss in Q2 should be > 1. In 2001, N.Dencker announced in [4] a positive answer to Q1, with answer 2 in Q2. His paper contains several new and interesting ideas; however, the author of this report was not able to understand thoroughly his article.

The Nirenberg-Treves conjecture is an important question of analysis, connecting a geometric (classical) property of symbols (Hamiltonians) to a priori inequalities for the quantized operators. The conventional wisdom on this problem turned out to be painfully wrong in the past, requiring the most careful examination of future claims.

References

[1] R.Beals, C.Fefferman, *On local solvability of linear partial differential equations*, Ann. of Math. **97** (1973), 482–498.
[2] N.Dencker, *The solvability of non-L^2-solvable operators*, Saint Jean de Monts meeting (1996).
[3] ———, *Estimates and solvability*, Arkiv.Mat. **37** (1999), 2, 221–243.
[4] ———, *On the sufficiency of condition (ψ)*, preprint (october 2001).
[5] C.Fefferman, D.H.Phong, *On positivity of pseudo-differential equations*, Proc. Nat. Acad. Sci. **75** (1978), 4673–4674.

[6] L.Hörmander, *On the theory of general partial differential operators*, Acta Math. **94** (1955), 161–248.
[7] ———, *Differential equations without solutions*, Math.Ann. **140** (1960), 169–173.
[8] ———, *Pseudo-differential operators and non-elliptic boundary value problems*, Ann. of Math. **83** (1966), 129–209.
[9] ———, *Propagation of singularities and semiglobal existence theorems for (pseudo-) differential operators of principal type*, Ann.of Math. **108** (1978), 569–609.
[10]———, *Pseudo-differential operators of principal type*, Singularities in boundary value problems, D.Reidel Publ.Co., Dortrecht, Boston, London, 1981.
[11]———, *The analysis of linear partial differential operators I–IV*, Springer Verlag, 1983–85.
[12]———, *Notions of convexity*, Birkhäuser, 1994.
[13]———, *On the solvability of pseudodifferential equations*, Structure of solutions of differential equations (M.Morimoto, T.Kawai, eds.), World Sci. Publishing, River Edge, NJ, 1996, 183–213.
[14]N.Lerner, *Sufficiency of condition* (ψ) *for local solvability in two dimensions*, Ann.of Math. **128** (1988), 243–258.
[15]———, *An iff solvability condition for the oblique derivative problem*, Séminaire EDP, Ecole Polytechnique (1990–91), exposé 18.
[16]———, *Nonsolvability in* L^2 *for a first order operator satisfying condition* (ψ), Ann.of Math. **139** (1994), 363–393.
[17]———, *When is a pseudo-differential equation solvable?*, Ann. Fourier **50** (2000), 2(spéc.cinq.), 443–460.
[18]———, *Energy methods via coherent states and advanced pseudo-differential calculus*, Multidimensional complex analysis and partial differential equations (, 177–201, eds.), AMS.
[19]———, *Perturbation and energy estimates*, Ann.Sci.ENS **31** (1998), 843–886.
[20]H.Lewy, *An example of a smooth linear partial differential equation without solution*, Ann.of Math. **66, 1** (1957), 155–158.
[21]S.Mizohata, *Solutions nulles et solutions non analytiques*, J.Math.Kyoto Univ. **1** (1962), 271–302.
[22]R.D.Moyer, *Local solvability in two dimensions: necessary conditions for the principal type case*, Mimeographed manuscript, University of Kansas (1978).
[23]L.Nirenberg, F.Treves, *Solvability of a first order linear partial differential equation*, Comm.Pure Appl.Math. **16** (1963), 331–351.
[24]———, *On local solvability of linear partial differential equations. I.Necessary conditions*, Comm.Pure Appl.Math. **23** (1970), 1–38.
[25]———, *On local solvability of linear partial differential equations. II.Sufficient conditions*, Comm.Pure Appl.Math. **23** (1970), 459–509.
[26]———, *On local solvability of linear partial differential equations. Correction*, Comm.Pure Appl.Math. **24** (1971), 279–288.
[27]J.-M.Trépreau, *Sur la résolubilité analytique microlocale des opérateurs pseudo-différentiels de type principal*, Thèse, Université de Reims (1984).

ICM 2002 · Vol. II · 721–732

Singular Integrals Meet Modulation Invariance

C. Thiele*

Abstract

Many concepts of Fourier analysis on Euclidean spaces rely on the specification of a frequency point. For example classical Littlewood Paley theory decomposes the spectrum of functions into annuli centered at the origin. In the presence of structures which are invariant under translation of the spectrum (modulation) these concepts need to be refined. This was first done by L. Carleson in his proof of almost everywhere convergence of Fourier series in 1966. The work of M. Lacey and the author in the 1990's on the bilinear Hilbert transform, a prototype of a modulation invariant singular integral, has revitalized the theme. It is now subject of active research which will be surveyed in the lecture. Most of the recent related work by the author is joint with C. Muscalu and T. Tao.

2000 Mathematics Subject Classification: 42B20, 47H60.
Keywords and Phrases: Fourier analysis, Singular integrals, Multilinear.

1. Multilinear singular integrals

A basic example for the notion of *singular integral* is a convolution operator

$$Tf(x) = K * f(x) = \int K(x-y)f(y)\,dy \tag{1.1}$$

whose convolution kernel K is not absolutely integrable. If K was absolutely integrable then we had trivially an a priori estimate

$$\|K * f\|_p \leq \|K\|_1 \|f\|_p \tag{1.2}$$

for $1 \leq p \leq \infty$. This follows by standard interpolation techniques from the two endpoints $p = 1, \infty$, which are true by trivial manipulations.

*Department of Mathematics, UCLA, Los Angeles, CA 90095-1555, USA. E-mail: thiele@math.ucla.edu

A basic point of singular integral theory is that an estimate of the form (1.2) may prevail for $1 < p < \infty$ with a constant $C_{p,K}$ instead of $\|K\|_1$ on the right hand side, if K is not absolutely integrable and the integral (1.1) is only defined in a distributional (principal value) sense. The most prominent example on the real line (indeed, all operators in this article will act on functions on the real line) is the Hilbert transform with $K(x) = 1/x$.

Taking formally Fourier transforms, one can write (1.1) as multiplier operator:

$$\widehat{Tf}(\xi) = \widehat{K}(\xi)\widehat{f}(\xi) =: m(\xi)\widehat{f}(\xi). \tag{1.3}$$

For the purpose of this survey a sufficiently interesting class of singular integrals is described in terms of the multiplier m by imposing the symbol estimates

$$(d/d\xi)^\alpha m(\xi) \leq C|\xi|^{-\alpha} \tag{1.4}$$

for $\alpha = 0, 1, 2$. We define the dual bilinear form

$$\Lambda(f_1, f_2) = \int (Tf_1(x))f_2(x)\,dx = \int_{\xi_1+\xi_2=0} \widehat{f_1}(\xi_1)\widehat{f_2}(\xi_2)m(\xi_1)\,d\sigma \tag{1.5}$$

where $d\sigma$ is the properly normalized Lebesgue measure on the hyperplane $\xi_1+\xi_2 = 0$. The natural generalization of estimate (1.2) using duality of L^p spaces then takes the form

$$|\Lambda(f_1, f_2)| \leq C_{p_1}\|f_1\|_{p_1}\|f_2\|_{p_2} \tag{1.6}$$

with $1/p_1 + 1/p_2 = 1$.

Estimate (1.6) can be related to square function estimates which are fundamental in singular integral theory. Let $(\psi_j)_{j\in\mathbb{Z}}$ be a family of functions such that $m_j := \widehat{\psi_j}$ is supported in the ball $B(0, 2^j)$ of radius 2^j around 0, vanishes on $B(0, 2^{j-2})$, and satisfies the symbol estimates (1.4) uniformly in j. By square function estimate we mean the inequality

$$\|(\sum_j |f * \psi_j|^2)^{1/2}\|_p \leq C_p\|f\|_p \tag{1.7}$$

which holds for $1 < p < \infty$. Now let m be any multiplier satisfying (1.4). It is easy to split it as $m(\xi_1) = \sum_j \widehat{\psi}_{1,j}(\xi_1)\widehat{\psi}_{2,j}(-\xi_1)$ for two families $\psi_{1,j}$ and $\psi_{2,j}$ as in the square function estimate. Then we have

$$|\Lambda(f_1, f_2)| = |\sum_j \int (f_1 * \psi_{1,j})(x)(f_2 * \psi_{2,j})(x)dx|.$$

Moving the sum inside the integral and applying Cauchy-Schwarz, Hölder, and (1.7) we obtain (1.6).

A natural generalization (see [14]) of (1.5) to multilinear forms is

$$\Lambda(f_1, \dots, f_n) = \int_{\xi_1+\dots+\xi_n=0} m(\xi_1, \dots, \xi_{n-1}) \prod_{j=1}^n \widehat{f_j}(\xi_j)\,d\sigma \tag{1.8}$$

with multipliers m satisfying

$$\partial^\alpha m(\xi') \le C|\xi'|^{-|\alpha|}. \tag{1.9}$$

Here $\xi' = (\xi_1, \dots, \xi_{n-1})$ and α runs through all multi- indices up to some order N. Note that the special role of the index n in the above is purely notational. The natural estimates to ask for are

$$|\Lambda(f_1, \dots, f_n)| \le C_{p_1,\dots,p_{n-1}} \prod_{j=1}^{n} \|f_j\|_{p_j} \tag{1.10}$$

for $1 < p_j < \infty$ with $\sum_j 1/p_j = 1$. In the special case that m is constant, $\Lambda(f_1, \dots, f_n)$ is a multiple of the integral of the pointwise product of the functions f_j and estimate (1.10) is simply Hölder's inequality.

We sketch a proof of (1.10). Without destroying the symbol estimates, we can split m into a finite sum of multipliers, each supported on a narrow cone with tip at the origin. Thus assume m is supported on such a cone consisting of rays having small angle with a vector η'.

We may assume by symmetry that ${\eta'}_1 = 1$ is the maximal component of η'. Then we can split m into pieces m_j satisfying (1.9) uniformly and supported in

$$(B(0, 2^j) \setminus B(0, 2^{j-2})) \times B(0, 2^{j+n})^{n-2}.$$

Introduce η_n such that $\sum_j \eta_j = 0$. By symmetry among the indices larger than 1 we may assume $\eta_2 \ge 1/n$. Then it is easy to arrange (see Figure "Cone") the support of m_j to be in

$$(B(0, 2^j) \setminus B(0, 2^{j-2})) \times (B(0, 2^{j+n}) \setminus B(0, 2^{j-n})) \times B(0, 2^{j+n})^{n-3}.$$

Using smoothness of the multiplier m_j we may use Fourier expansion to write it as rapidly converging sum of multipliers of elementary tensor form

$$\widehat{\psi}_{1,j}(\xi_1)\widehat{\psi}_{2,j}(\xi_2) \prod_{l=3}^{n} \widehat{\phi}_{l,j}(\xi_l)$$

with $\xi_n = -\sum_{j=1}^{n-1} \xi_{n-1}$. The symbol estimates prevail for these elementary tensors, and thus we observe

$$(d/d\xi)^\alpha(\phi_{l,j})(\xi) \le C2^{-\alpha j} \tag{1.11}$$

for all derivatives up to order N. Observe that $\widehat{\psi}_{l,j}$ are essentially as in (1.7), and $\widehat{\phi}_{l,j}$ are similar but fail to be supported away form the origin. Applying the elementary tensor multiplier form to $f_1, \dots, f_2$ is the same as applying a constant multiplier to $\psi_{1,j} * f_1, \dots, \phi_{n,j} * f_n$. Estimate (1.10) then follows from

$$\sum_j \int \prod_{l=1}^{2} (\psi_{l,j} * f_l)(x) \prod_{l=3}^{n} \phi_{l,j} * f_l(x)\, d\sigma$$

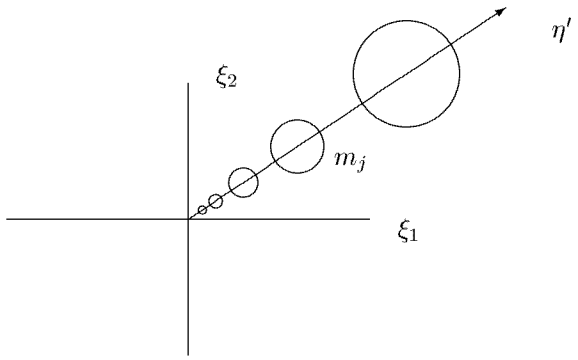


Figure 1: "Cone"

$$\leq C \prod_{l=1}^{2} \|(\sum_j |f_l * \psi_{l,j}|^2)^{1/2}\|_{p_l} \prod_{l=3}^{\infty} \|\sup_j |f_l * \phi_{l,j}|\|_{p_l} \leq C \prod_{l=1}^{n} \|f_l\|_{L^p_l}.$$

Here we have used for $l = 1, 2$ the square function estimate (1.7) and for $l > 2$ the equally fundamental Hardy Littlewood maximal inequality

$$\|\sup_j |f * \phi_{l,j}|\|_{L^p} \leq C_p \|f\|_p$$

which is valid due to (1.11).

2. Modulation invariance

Modulation M_η with parameter $\eta \in \mathbb{R}$ is defined to be multiplication by a character:

$$M_\eta f(x) := f(x) e^{2\pi i \eta x}.$$

This amounts to a translation of the Fourier transform of f.

We shall be interested in multilinear forms Λ which have modulation symmetries in the sense

$$\Lambda(f_1, \dots, f_n) = \Lambda(M_{\eta_1} f_1, \dots M_{\eta_n} f_n) \tag{2.1}$$

for all vectors $\eta = (\eta_1, \dots, \eta_n)$ in a subspace Γ of the hyperplane given by $\sum \eta_j = 0$.

If Λ is given in multiplier form (1.8), then (2.1) is equivalent to a translation symmetry of the multiplier m:

$$m(\xi_1, \dots, \xi_n) = m(\xi_1 + \eta_1, \dots, \xi_n + \eta_n). \tag{2.2}$$

Such a symmetry with nontrivial η is inconsistent with the symbol estimates (1.9) unless m is constant. Namely, by iterating (2.2), any point with nonvanishing

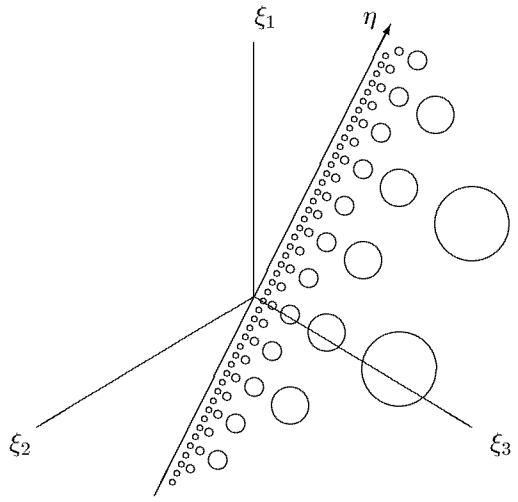


Figure 2: "Circles"

derivative of m can by translated to a point far away from the origin, until the value of the derivative, which remains constant at the translated points, contradicts (1.9).

A natural replacement for (1.9) in the presence of modulation symmetry along vectors in Γ has been introduced by Gilbert/Nahmod [6]:

$$\partial^\alpha m(\xi') \leq C\mathrm{dist}(\xi', \Gamma')^{-|\alpha|}. \tag{2.3}$$

Here Γ' is the projection of Γ onto the first $n-1$ coordinates. Figure "Circles" indicates the regions in which multipliers of the form (2.3) can be thought of as being essentially constant.

The following theorem is due to [6] in the case $n = 3$ and to [16] in general:

Theorem 2.1 *Assume $k := \dim(\Gamma) < n/2$, and assume that Γ is non-degenerate in the sense that for any $1 \leq i_1 < \cdots < i_k \leq n$ the space Γ is the graph of a function in the variables $\xi_{i_1}, \ldots, \xi_{i_k}$. Assume m satisfies (2.3). Then Λ as in (1.8) satisfies (1.10) whenever $\sum 1/p_j = 1$ and $1 < p_j \leq \infty$ for all p_j.*

We remark that it is unknown whether the condition $\dim(\Gamma) < n/2$ can be relaxed in this theorem.

The forms Λ have dual multilinear operators. Theorem 2.1 implies a priori estimates for these multilinear operators. Moreover, these multilinear operators satisfy estimates which cannot be formulated in terms of L^p estimates for Λ. Let $(p_1, \ldots, p_n)$ be a tuple of real numbers or ∞ such that at most one of these numbers

is negative. If all of them are nonnegative, we say Λ is of type $(p_1,\dots,p_n)$ if (1.10) holds. If one of them, say p_j, is negative, then we define the dual operator T by

$$\Lambda(f_1,\dots,f_n)=\int T(f_1,\dots,f_{j-1},f_{j+1},\dots,f_n)(x)f_j(x)\,dx.$$

We then say that Λ is of type $(p_1,\dots,p_n)$ if

$$\|T(f_1,\dots,f_{j-1},f_{j+1},\dots,f_n)\|_{p_j'}\le C\prod_{i\neq j}\|f_i\|_{p_i}$$

where $p_j'=p_j/(p_j-1)$. Observe $0<p_j'<1$. The following theorem is again due to [6] ($n=3$) and [16]:

Theorem 2.2 *Let Γ and Λ be as in Theorem 2.1. Then Λ is of type $(p_1,\dots,p_n)$ if $\sum_j 1/p_j=1$, at most one of the p_j is negative, none of the p_j is in $[0,1]$, and*

$$1/p_{i_1}+\cdots+1/p_{i_r}<\frac{n-2\dim(\Gamma)+r}{2}$$

for all $1\le i_1<\cdots<i_r\le n$ and $1\le r\le n$.

A basic example of a modulation invariant form Λ is when $n=3$ and $m(\xi_1,\xi_2)$ is constant on both sides of a line Γ but not globally constant. With proper choice of constants this form can be written as

$$\Lambda_\alpha(f_1,f_2,f_3)=\int B_\alpha(f_1,f_2)(x)f_3(x)\,dx$$

with the bilinear Hilbert transform

$$B_\alpha=p.v.\int f_1(x-t)f_2(x-\alpha t)\frac{1}{t}\,dt$$

and a (projective) parameter α determining the direction of the line Γ. Theorems 2.1 and 2.2 in this special case are due to [10] and [11].

For the bilinear Hilbert transform nondegeneracy specializes to the condition $\alpha\notin\{0,1,\infty\}$, and the conclusion of both theorems can be summarized to

$$\|B_\alpha(f_1,f_2)\|_p\le C_{p_1,p_2}\|f_1\|_{p_1}\|f_2\|_{p_2}\tag{2.4}$$

provided $1<p_1,p_2\le\infty$ and $2/3<p<\infty$. The set of types of such Λ_α is the convex hull of the open triangles a,b,d in Figure "Hexagon" which depicts the plane of $(1/p_1,1/p_2,1/p_3)$ with $\sum_j 1/p_j=1$. It is unknown whether the type-region of Λ_α extends to the open triangle e and its symmetric counterparts.

We point out a related result by M. Lacey [9]:

Theorem 2.3 *The maximal truncations of the bilinear Hilbert transform,*

$$B_\alpha^{\max}(f,g)(x):=\sup_{\epsilon>0}\left|\int_{\mathbb{R}\setminus[-\epsilon,\epsilon]}f(x-t)g(x-\alpha t)\frac{1}{t}\,dt\right|$$

also satisfy (2.4) provided α is not degenerate.

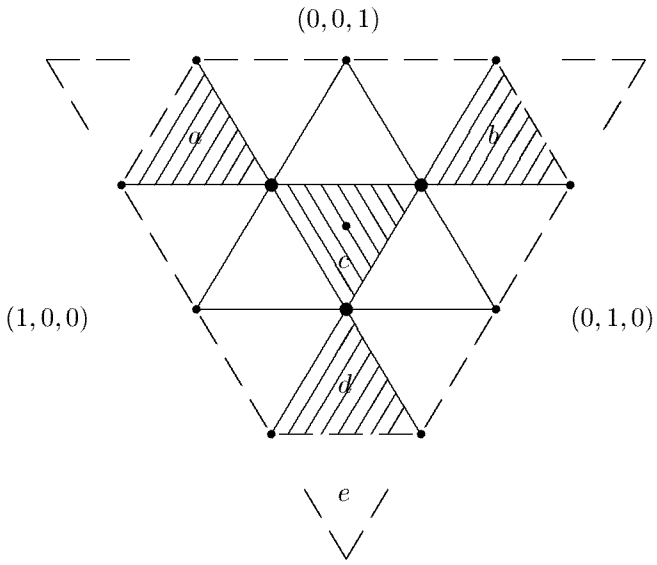


Figure 3: "Hexagon"

This is stronger than the bounds for the bilinear Hilbert transform itself.

The main difference in proving the theorems in this section compared to the discussion in Section is that it is not sufficient to split the functions f_k into frequency parts supported in $B(0,2^j) \setminus B(0,2^{j-2})$. The special role that is attributed to the zero frequency by this splitting is obsolete in the modulation invariant setting. Instead one has to consider frequency bands of f_k away from the origin and very narrow, such as intervals $[N-\epsilon, N+\epsilon]$ for large N and small ϵ. Geometrically these bands can be viewed as the projections of the circles in Figure "Circles" onto the projected coordinate axes. Handling thin frequency bands requires a new set of techniques. Prior to the work [10] and [11] these techniques have been pioneered in [2] and [5] where the Carleson operator

$$Cf(x) = \sup_{\xi} |p.v. \int e^{iy\xi} f(x-y) \frac{1}{y}\, dy|$$

has been estimated. Note that this operator is modulation invariant, $C(f) = C(M_\eta f)$. See also [12]. Most theorems discussed in this survey have a simpler but significant model theorem in the dyadic setting, see for example [17], [22],

3. Uniform estimates

Theorem 2.1 excludes certain degenerate subspaces Γ. For some degenerate Γ the multilinear forms split into simpler objects and one can provide L^p estimates

also in these degenerate cases; we will give examples below. This raises the question whether one can prove bounds on Λ uniformly in the choice of Γ, as Γ approaches one of these degenerate cases.

Substantial progress on this question has only been made in the case $\dim(\Gamma) = 1$.

Theorem 3.1 *Let $n \geq 3$ and $(\eta_1, \dots, \eta_n)$ be a unit vector spanning the space Γ, and assume $\eta_j \neq 0$ for all j. Define the metric*

$$d(x,y) := \sup_{1 \leq j \leq n} \frac{|x_i - y_i|}{|\eta_i|}$$

and write $d(x,\Gamma) := \inf_{y \in \Gamma} d(x,y)$. Suppose m satisfies the estimate

$$\partial_{\eta'}^{\alpha} m(\eta') \leq \prod_{j=1}^{n} (\eta_j d(\eta, \Gamma))^{-\alpha_j} \tag{3.1}$$

for all partial derivatives $\partial_{\eta'}^{\alpha}$ up to order N. Then (1.10) holds for all $2 < p_j < \infty$ with $\sum_j 1/p_j = 1$ with the bounds uniform in the choice of Γ.

We discuss uniform estimates for the special case of the bilinear Hilbert transform. The degenerate directions for Γ occur when the vector η is perpendicular to one of the three projected coordinate axes (see Figure "Circles"). One of the degenerate cases ($\alpha = 1$) gives rise to the operator

$$B_1(f_1, f_2) = H(f_1 \cdot f_2)$$

(Hilbert transform of the pointwise product) or its dual operators

$$f_2 \cdot H(f_3) \ , \quad f_1 \cdot H(f_3).$$

Besides the usual homogeneity $\sum_j 1/p_j = 1$, the only constraint for these operators to be of type (p_1, p_2, p_3) is $1 < p_3 < \infty$. In Figure "Hexagon" this region is the strip bounded by the horizontal lines through $(0,0,1)$ and $(1,0,0)$.

Thus one expects the constants in the L^p estimates to be uniform as α approaches 1 in the intersection of this strip and the convex hull of triangles a, b, d. The above theorem provides uniform estimates in the inner triangle c. This special case of Theorem 3.1 was previously shown by Grafakos/Li [7], and Li [13] has shown uniform estimates in triangles a and b. These results together give uniform bounds in the convex hull of a, b, c. Uniform estimates near the points $(1,0,0)$ and $(0,1,0)$ remain an open question. Prior to the work of Grafakos/Li [7], weak type uniform bounds were shown [23], [24] in the common boundary point of triangles a and c (and by symmetry also b and c).

The multiplier condition (3.1) gives essentially constant multipliers on regions adapted to the slope of Γ, see Figure "Ellipses". Observe that all ellipses at a given scale project essentially onto disjoint regions when projected to any one one of the coordinate axes. Handling these adapted regions uniformly requires considerable refinements of the arguments in [10] and [11].

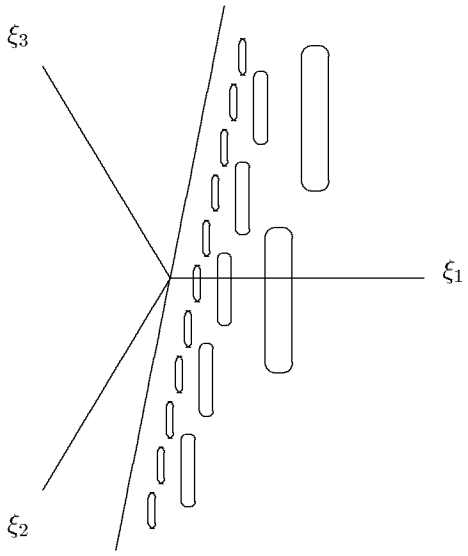


Figure 4: "Ellipses"

We mention that closely related to the topic of uniform estimates for the bilinear Hilbert transform is that of bilinear multiplier estimates for multipliers which are singular along a curve rather than a line, provided the curve is tangent to a degenerate direction. Results for such multipliers have been found by Muscalu [15] and Grafakos/Li [8].

We conclude this section with a remark on the history of the bilinear Hilbert transform. Calderon is said to have considered the bilinear Hilbert transform in the 1960's while studying what has been named Calderon's first commutator. This is the bilinear operator

$$\mathcal{C}(A,f)(x) = p.v. \int \frac{A(x)-A(y)}{(x-y)^2} f(y)\, dy.$$

It can be viewed as a bilinear operator in the derivative A' of A and the function f, and as such has a multplier form as in (1.8). To see this, we can write $C(A,f)$ in terms of A' as a superposition of bilinear Hilbert transforms:

$$\begin{aligned}\mathcal{C}(A,f)(x) &= p.v. \int \int_0^1 A'(x+\alpha(y-x)) \frac{1}{x-y} f(y)\, d\alpha dy \\ &= \int_0^1 B_\alpha(f,A')(x)\, d\alpha.\end{aligned}$$

The estimate Calderon was looking for was

$$\|C(A,f)\|_2 \leq \|A'\|_\infty \|f\|_2. \tag{3.2}$$

Thus he needed good control over the constant C_α as α approaches 0 or 1. However, even finiteness of C_α was not known to Calderon. Sufficiently good control over C_α was first established in [23].

The multiplier of $\mathcal{C}(A', f)$ is more regular than that of the bilinear Hilbert transform, and Calderon, quitting his attempts to estimate the bilinear Hilbert transform, proved estimate (3.2) by refinements of the methods in Section (see [1]).

4. More multilinear operators

Theorem 2.1 discusses multipliers singular at a single subspace Γ'. Cut and paste arguments easily allow to generalize the theorem to the case of multipliers singular at finitely many subspaces $\Gamma_1', \dots, \Gamma_k'$, provided each subspace satisfies the dimension and non-degeneracy conditions of Theorem 2.1.

Interesting phenomena occur for multipliers singular at several subspaces Γ_1', $\dots, \Gamma_k'$ which do not satisfy the conditions of Theorem 2.1. Some operators corresponding to multipliers singular at degenerate subspaces can be written in terms of pointwise products and lower degree operators and thus can be trivially shown to satisfy L^p estimates. If m is singular at several such subspaces, the trivial splitting may no longer be possible, and one has to do a much more subtle analysis.

We consider the special case when the spaces $\Gamma_1', \dots, \Gamma_k'$ are hyperplanes and the multiplier is the characteristic function of one of the infinite simplices been cut out of $\mathbb{R}^n$ by these hyperplanes, see Figure "Wedge". A basic example is the trilinear operator

$$T(f_1, f_2, f_3)(x) = \int_{\alpha_1\xi_1 < \alpha_2\xi_2 < \alpha_3\xi_3} \prod_{j=1}^{3} \widehat{f}_j(\xi_j) e^{2\pi i x \xi_j} \, d\xi_j$$

and its associated fourlinear form

$$\Lambda(f_1, f_2, f_3, f_4) = \int_{\sum_{j=1}^{4} \xi_j = 0, \alpha_1\xi_1 < \alpha_2\xi_2 < \alpha_3\xi_3} \prod_{j=1}^{4} \widehat{f}_j(\xi_j) \, d\sigma. \tag{4.1}$$

Here $\alpha_1, \alpha_2, \alpha_3$ are real parameters. If we had only one of the two constraints $\alpha_1\xi_1 < \alpha_2\xi_2$ or $\alpha_2\xi_2 < \alpha_3\xi_3$, then these operators would decompose trivially.

There is a Zariski open set of values of $(\alpha_1, \alpha_2, \alpha_3)$ for which Λ and T are well behaved. The following theorem proved in [18] states such estimates for the generic point $(1, 1, 1)$.

Theorem 4.1 *For $\alpha_1, \alpha_2, \alpha_3 = 1$ the form Λ as in (4.1) satisfies estimates*

$$\Lambda(f_1, f_2, f_3, f_4) \leq C_{p_1, \dots, p_4} \prod_{j=1}^{4} \|f_j\|_{p_j}$$

if $1 < p_j < \infty$ and $\sum_j 1/p_j = 1$. The trilinear form T satisfies in addition estimates mapping into L^p with $p < 1$, in particular

$$\|T(f_1, f_2, f_3)\|_{2/3} \leq C \prod_{j=1}^{3} \|f_j\|_2.$$

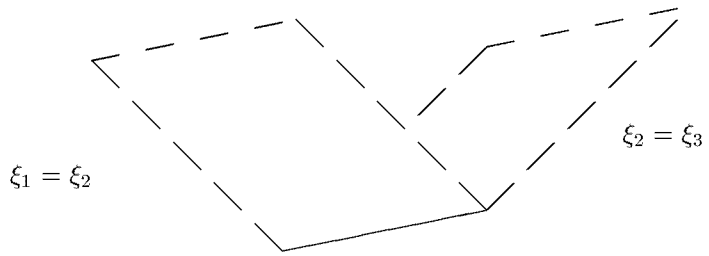


Figure 5: "Wedge"

An example for a degenerate choice of $(\alpha_1, \alpha_2, \alpha_3)$ is $(1, -1, 1)$. In this case there is a negative result [19]:

Theorem 4.2 *For $\alpha_1 = 1, \alpha_2 = -1, \alpha_3 = 1$ the a priori estimate*

$$\|T(f_1, f_2, f_3)\|_{2/3} \leq C \prod_{j=1}^{3} \|f_j\|_2$$

does not hold.

Theorem 4.2 is proved by applying T to functions f_1, f_2, f_3 which are suitable truncations of imaginary Gaussians (chirps) $e^{i\beta x^2}$. The operator of Theorem 4.2 appears naturally in eigenfunction expansions of one dimensional Schrödinger operators, see the work of Christ/Kiselev [3],[4]. A positive result on discrete models of these expansions using the modulation invariant theory can be found in [20].

References

[1] Calderon A. P., *Commutators of singular integral operators.* Proc. Natl. Acad. Sci. USA, Vol. 53, 1092–1099. [1977]

[2] Carleson L., *On convergence and growth of partial sums of Fourier series.* Acta Math. 116, 135–157. [1966]

[3] Christ, M., Kiselev, A., *WKB asymptotics of generalized eigenfunctions of one-dimensional Schrödinger operators*, J. Funct. An. **179**, no. 2, 426–447. [2001]

[4] Christ, M., Kiselev, A., *WKB and spectral analysis of one-dimensional Schrödinger operators with slowly varying potential*, Comm. Math. Phys. **218**, 245–262. [2001]

[5] Fefferman C., *Pointwise convergence of Fourier series.* Ann. Math. 98, 551–571. [1973]

[6] Gilbert J., Nahmod A., *Boundedness of bilinear operators with non-smooth symbols* Math. Res. Lett. **7**, 767–778. [2000]
[7] Grafakos L., Li X, *Uniform bounds for the bilinear Hilbert transform I*, preprint. [2000]
[8] Grafakos L., Li X, *The disc as multiplier* preprint. [2000]
[9] M. Lacey, *The bilinear maximal function maps into L^p for $2/3 < p \leq 1$* Ann. Math (2) **151** (2000) no. 1, 35–57.
[10] Lacey M., Thiele C., *L^p estimates on the bilinear Hilbert transform for $2 < p < \infty$.* Ann. Math. 146, 693–724. [1997]
[11] Lacey M., Thiele C., *On Calderon's conjecture.* Ann. Math (2) **149** (1999) no. 2, 475–496.
[12] Lacey M., Thiele C., *A proof of boundedness of the Carleson operator.* Math. Res. Lett. *7* [2000] no. 4, 361–370.
[13] Li X, *Uniform bounds for the bilinear Hilbert transform II*, preprint. [2000]
[14] Meyer Y., Coifman R. R., *Opérateurs multilinéaire*, Hermann, Paris, [1991]
[15] Muscalu C. *L^p estimates for multipliers given by singular symbols.* PhD Thesis, Brown University [2000]
[16] Muscalu C., Tao T., Thiele C., *Multilinear operators given by singular symbols*, to appear in J. Amer. Math. Soc.
[17] Muscalu C., Tao T., Thiele C., *L^p estimates for the biest I. The Walsh case*, preprint. [2001]
[18] Muscalu C., Tao T., Thiele C., *L^p estimates for the biest II. The Fourier case*, preprint. [2001]
[19] Muscalu C., Tao T., Thiele C., *A counterexample to a multilinear endpoint question of Christ and Kiselev* to appear in Math. Res. Lett.
[20] Muscalu C., Tao T., Thiele C., *A Carleson type theorem for a Cantor group model of the scattering transform*, preprint.
[21] Thiele C., Ph. D. Thesis, Yale University. [1995]
[22] Thiele C., *The quartile operator and pointwise convergence of Walsh series*, Trans. Amer. Math. Soc. **352**, [2000] (no. 12), 5745–5766.
[23] Thiele C., *On the Bilinear Hilbert transform.* Universität Kiel, Habilitationsschrift. [1998]
[24] Thiele C., *A uniform estimate.* Ann. Math. **157**, 1–45. [2002]

ICM 2002 · Vol. II · 733–742

Asymptotics of Polynomials and Eigenfunctions

S. Zelditch*

Abstract

We review some recent results on asymptotic properties of polynomials of large degree, of general holomorphic sections of high powers of positive line bundles over Kähler manifolds, and of Laplace eigenfunctions of large eigenvalue on compact Riemannian manifolds. We describe statistical patterns in the zeros, critical points and L^p norms of random polynomials and holomorphic sections, and the influence of the Newton polytope on these patterns. For eigenfunctions, we discuss L^p norms and mass concentration of individual eigenfunctions and their relation to dynamics of the geodesic flow.

2000 Mathematics Subject Classification: 35P20, 30C15, 32A25, 58J40, 60D05, 81S10, 14M25.
Keywords and Phrases: Random polynomial, Holomorphic section of positive line bundle, Distribution of zeros, Correlation between zeros, Bergman-Szego kernels, Newton polytope, Laplace eigenfunction, Spectral projections, L^p-norms, Quantum ergodicity.

1. Introduction

In many measures of 'complexity', eigenfunctions $\sqrt{\Delta}\varphi_\lambda = \lambda\varphi_\lambda$ of first order elliptic operators behave like polynomials $p(x) = \sum_{|\alpha|\leq x} c_\alpha x^\alpha$ of degree $N \sim \lambda$ [6]. The basic example we have in mind is the Laplacian Δ on a compact Riemannian manifold (M, g), but the same is true of Schroedinger operators. The comparison is more than an analogy, since polynomials of degree N are eigenfunctions of a first order elliptic system.

The comparison between eigenfunctions and polynomials is an essentially local one, most accurate on small balls $B(x_0, \frac{1}{\lambda})$. Globally, eigenfunctions reflect the dynamics of the geodesic flow G^t on the unit (co-)tangent bundle S^*M. This is one of the principal themes of quantum chaos.

*Department of Mathematics, Johns Hopkins University, Baltimore, Maryland 21218, USA. E-mail: zelditch@math.jhu.edu

In this article, we review some recent results on the asymptotics of polynomials and eigenfunctions, concentrating on our work in collaboration with P. Bleher, A. Hassell, B. Shiffman, C. Sogge, J. Toth and M. Zworski. A unifying feature is the asymptotic properties of reproducing kernels, namely Szegö kernels $\Pi_N(z,w)$ in the case of polynomials, and spectral projections $E_\lambda(x,y)$ for intervals $[\lambda,\lambda+1]$ in the case of eigenfunctions of $\sqrt{\Delta}$. For other recent expository articles, see [9, 26].

2. Polynomials

There are several sources of interest in random polynomials. One is the desire to understand typical properties of real and complex algebraic varieties, and how they depend on the coefficients of the defining equations. Another is their use as a model for the local behavior of more general eigenfunctions. A third is that they may be viewed as the eigenvectors of random matrices. Just as random matrices model the spectra of 'quantum chaotic' systems, so random polynomials model their eigenfunctions.

2.1. $SU(m+1)$ polynomials on $\mathbb{CP}^m$ and holomorphic sections

Complex polynomials of degree $\leq p$ in m variables form the vector space

$$\mathcal{P}_p^m := \{f(z_1,\dots,z_m) = \sum_{\alpha\in\mathbb{N}^m:|\alpha|\leq p} c_\alpha z_1^{\alpha_1}\cdots z_m^{\alpha_m},\ \ c_\alpha\in\mathbb{C}\}\ .$$

To put a probability measure on $\mathcal{P}_p^m$ is to regard the coefficients c_α as random variables. The simplest measures are Gaussian measures corresponding to inner products on $\mathcal{P}_p^m$. By homogenizing f to $F(z_0,z_1,\cdots,z_m)$ of degree p, we may identify $\mathcal{P}_p^m$ with the space $H^0(\mathbb{CP}^m,\mathcal{O}(p))$ of holomorphic sections of the pth power of the hyperplane bundle. It carries the standard $\mathrm{SU}(m+1)$-invariant Fubini-Study inner product $\langle F_1,\bar{F}_2\rangle_{FS} = \int_{S^{2m+1}} F_1\bar{F}_2\,d\sigma$, where $d\sigma$ is Haar measure on the $(2m+1)$-sphere S^{2m+1}. An orthonormal basis of $H^0(\mathbb{CP}^m,\mathcal{O}(p))$ is given by $\{\frac{z^\alpha}{||z^\alpha||_{FS}}\}$. The corresponding $\mathrm{SU}(m+1)$-invariant Gaussian measure γ_δ is defined by

$$d\gamma_\delta(s) = \frac{1}{\pi^{k_p}}e^{-|\lambda|^2}d\lambda,\ \ s=\sum_{|\alpha|\leq p}\lambda_\alpha\frac{z^\alpha}{||z^\alpha||_{FS}}.$$

Thus, the coefficients λ_α are independent complex Gaussian random variables with mean zero and variance one.

More generally, we can define Gaussian ensembles of holomorphic sections $H^0(M,L^N)$ of powers of a positive line bundle over any Kähler manifold (M,ω). Endowing L with the unique hermitian metric h of curvature form ω, we induce an inner product $\langle,\rangle$ on $H^0(M,L^N)$ and a Gaussian measure γ_N. We denote the unit sphere in $H^0(M,L^N)$ relative to this inner product by $SH^0(M,L^N)$. The Haar measure on $SH^0(M,L^N)$ will be denoted μ_N. It is closely related to the Gaussian measure.

2.2. Zeros

The problems we discuss in this section involve the geometry of zeros of sections $s \in H^0(M, L^N)$ of general positive line bundles. There is a similar story for critical points.

- **Problem 1** How are the simultaneous zeros $Z_s = \{z : s_1(z) = \cdots = s_k(z) = 0\}$ of a k-tuple $s = (s_1, \ldots, s_k)$ of typical holomorphic sections distributed?
- **Problem 2** How are the zeros correlated? When $k = m$, the simultaneous zeros form a discrete set. Do zeros repel each other like charged particles? Or behave independently like particles of an ideal gas? Or attract like gravitating particles?

By the distribution of zeros we mean either the current of integration over Z_s or more simply the Riemannian $(2m-2k)$-volume measure $(|Z_s|, \varphi) = \int_{Z_s} \varphi d\mathrm{Vol}_{2m-2k}$. By the n-point zero correlation functions, we mean the generalized functions

$$K_{nk}^N(z^1, \ldots, z^n)dz = \mathbf{E}|Z_s|^n,$$

where $|Z_s|^n$ denotes the product of the measures $|Z_s|$ on the punctured product $M_n = \{(z^1, \ldots, z^n) \in M \times \cdots \times M : z^p \neq z^q \text{ for } p \neq q\}$ and where dz denotes the product volume form on M_n.

The answer to Problem 1 is that zeros almost surely become uniformly distributed relative to the curvature ω of the line bundle [18]. Curvature causes sections to oscillate more rapidly and hence to vanish more often. More precisely, we consider the space $\mathcal{S} = \prod_{N=1}^{\infty} SH^0(M, L^N)$ of random sequences, equipped with the product measure measure $\mu = \prod_{N=1}^{\infty} \mu_N$. An element in $\mathcal{S}$ will be denoted $\mathbf{s} = \{s_N\}$. Then, $\frac{1}{N} Z_s \to \omega$, as $N \to \infty$ for almost every $\mathbf{s}$.

The answer to Problem 2 is more subtle: it depends on the dimension. We assume $k = m$ so that almost surely the simultaneous zeros of the k-tuple of sections form a discrete set. We find that these zeros behave almost independently if they are of distance $\geq \frac{D}{\sqrt{N}}$ apart for $D \gg 1$. So they only interact on distance scales of size $\frac{1}{\sqrt{N}}$. Since also the density of zeros in a unit ball $B_1(z_0)$ around z_0 grows like N^m, we rescale the zeros in the $1/\sqrt{N}$-ball $B_{1/\sqrt{N}}(z_0)$ by a factor of $\sqrt{N}$ to get configurations of zeros with a constant density as $N \to \infty$. We thus rescale the correlation functions and take the *scaling limits*

$$\widetilde{K}_{nkm}^{\infty}(z^1, \ldots, z^n) = \lim_{N \to \infty} K_{1k}^N(z_0)^{-n} K_{nk}^N(z_0 + \frac{z^1}{\sqrt{N}}, \ldots, z_0 + \frac{z^n}{\sqrt{N}}). \tag{2.1}$$

In [1], we proved that the scaling limits of these correlation functions were universal, i.e. independent of M, L, ω, h. They depend only on the dimension m of the manifold and the codimension k of the zero set.

In [2], we found explicit formulae for these universal scaling limits. In the case $n = 2$, $\widetilde{K}_{2km}^{\infty}(z^1, z^2)$, depends only on the distance between the points z^1, z^2, since it is universal and hence invariant under rigid motions. Hence it may be written as:

$$\widetilde{K}_{2km}^{\infty}(z^1, z^2) = \kappa_{km}(|z^1 - z^2|). \tag{2.2}$$

We refer to [1] for details.

Theorem 1 [2] *The pair correlation functions of zeros when $k = m$ are given by*

$$\kappa_{mm}(r) = \begin{cases} \frac{m+1}{4} r^{4-2m} + O(r^{8-2m}), & as \ \ r \to 0 \\ 1 + O(e^{-Cr^2}), \ \ (C > 0) & as \ r \to \infty. \end{cases} \tag{2.3}$$

When $m = 1, \kappa_{mm}(r) \to 0$ as $r \to 0$ and one has "zero repulsion." When $m = 2$, $\kappa_{mm}(r) \to 3/4$ as $r \to 0$ and one has a kind of neutrality. With $m \geq 3$, $\kappa_{mm}(r) \nearrow \infty$ as $r \to 0$ and there is some kind of attraction between zeros. More precisely, in dimensions greater than 2, one is more likely to find a zero at a small distance r from another zero than at a small distance r from a given point; i.e., zeros tend to clump together in high dimensions.

One can understand this dimensional dependence heuristically in terms of the geometry of the discriminant varieties $\mathcal{D}_N^m \subset H^0(M, L^N)^m$ of systems $S = (s_1, \ldots, s_m)$ of m sections with a 'double zero'. The 'separation number' $sep(F)$ of a system is the minimal distance between a pair of its zeros. Since the nearest element of $\mathcal{D}_N^m$ to F is likely to have a simple double zero, one expects: $sep(F) \sim \sqrt{dist(F, \mathcal{D}_N^m)}$. Now,the degree of $\mathcal{D}_N^m$ is approximately N^m. Hence, the tube $(\mathcal{D}_N^m)_\epsilon$ of radius ϵ contains a volume $\sim \epsilon^2 N^m$. When $\epsilon \sim N^{-m/2}$, the tube should cover $PH^0(M, L^N)$. Hence, any section should have a pair of zeros whose separation is $\sim N^{-m/4}$ apart. It is clear that this separation is larger than, equal to or less than $N^{-1/2}$ accordingly as $m = 1, m = 2, m \geq 3$.

2.3. Bergman-Szegö kernels

A key object in the proof of these results is the Bergman-Szegö kernel $\Pi_N(x, y)$, i.e. the kernel of the orthogonal projection onto $H^0(M, L^N)$ with respect to the Kähler form ω. For instance, the expected distribution of zeros is given by $\mathbf{E}_N(Z_f) = \frac{\sqrt{-1}}{2\pi} \partial\bar{\partial} \log \Pi_N(z, z) + \omega$. Of even greater use is the joint probability distribution (JPD) $D_N(x^1, \ldots, x^n; \xi^1, \ldots, \xi^n; z^1, \ldots, z^n)$ of the random variables $x^j(s) = s(z^j)$, $\xi^j(s) = \nabla s(z^j)$, which may be expressed in terms of Π_N and its derivatives. In turn, the correlation functions may be expressed in terms of the JPD by $K^N(z^1, \ldots, z^n) = \int D_N(0, \xi, z) \prod_{j=1}^n \left(\|\xi^j\|^2 d\xi^j \right) d\xi$ [1].

The scaling asymptotics of the correlation functions then reduce to scaling asymptotics of the Bergman-Szegö kernel: In normal coordinates $\{z_j\}$ at $P_0 \in M$ and in a 'preferred' local frame for L, we have [1]:

$$\begin{aligned} & \frac{\pi^m}{N^m} \Pi_N(P_0 + \frac{u}{\sqrt{N}}, \frac{\theta}{N}; P_0 + \frac{v}{\sqrt{N}}, \frac{\varphi}{N}) \\ \sim \ & e^{i(\theta - \varphi) + u \cdot \bar{v} - \frac{1}{2}(|u|^2 + |v|^2)} \left[1 + b_1(u, v) N^{-\frac{1}{2}} + \cdots \right]. \end{aligned}$$

To be precise, Π_N is the natural lift of the kernel as an equivariant kernel on the boundary ∂D^* of the unit (co-) disc bundle of L^*. Note that $e^{i(\theta - \varphi) + u \cdot \bar{v} - \frac{1}{2}(|u|^2 + |v|^2)}$ is the Bergman-Szegö kernel of the Heisenberg group. These asymptotics use the Boutet de Monvel -Sjostrand parametrix for the Bergman-Szegö kernel [4], as applied in [29] to the Fourier coefficients of the kernel on powers of positive line bundles.

2.4. Polynomials with fixed Newton polytope

The well-known Bernstein-Kouchnirenko theorem states that the number of simultaneous zeros of (a generic family of) m polynomials with Newton polytope P equals $m!Vol(P)$. Recall that the Newton polytope P_f of a polynomial is the convex hull of its support $S_f = \{\alpha \in \mathbb{Z}^m : c_\alpha \neq 0\}$. Using the homogenization map $f \to F$, the space of polynomials f whose Newton polytope P_f contained in P may be identified with a subspace

$$H^0(\mathbb{CP}^m, \mathcal{O}(p), P) = \{F \in H^0(\mathbb{CP}^m, \mathcal{O}(p)) : P_f \subset P\} \tag{2.4}$$

of $H^0(\mathbb{CP}^m, \mathcal{O}(p))$.

The problem we address in this section is:

- **Problem 3** How does the Newton polytope influence on the distribution of zeros of polynomials?

Again, one could ask the same question about L^2 mass, critical points and so on and obtain a similar story. In [19] we explore this influence in a statistical and asymptotic sense. The main theme is that for each property of polynomials under study, P gives rise to *classically allowed regions* where the behavior is the same as if no condition were placed on the polynomials, and *classically forbidden regions* where the behavior is exotic.

Let us define these terms. If $P \subset \mathbb{R}^m_+$ is a convex integral polytope, then the *classically allowed region* for polynomials in $H^0(\mathbb{CP}^m, \mathcal{O}(p), P)$ is the set

$$\mathcal{A}_P := \mu_\Sigma^{-1}\left(\frac{1}{p}P^\circ\right) \subset \mathbb{C}^{*m}$$

(where P° denotes the interior of P), and the *classically forbidden region* is its complement $\mathbb{C}^{*m} \setminus \mathcal{A}_P$. Here, $\mu_\Sigma(z) = \left(\frac{|z_1|^2}{1+\|z\|^2}, \ldots, \frac{|z_m|^2}{1+\|z\|^2}\right)$ is the moment map of $\mathbb{CP}^m$.

The result alluded to above is statistical. Since we view the polytope P of degree p as placing a condition on the Gaussian ensemble of $SU(P)$ polynomials of degree p, we endow $H^0(\mathbb{CP}^m, \mathcal{O}(p), P)$ with the *conditional probability measure* $\gamma_\delta|_P$:

$$d\gamma_{\delta|P}(s) = \frac{1}{\pi^{\#P}} e^{-|\lambda|^2} d\lambda, \quad s = \sum_{\alpha \in P} \lambda_\alpha \frac{z^\alpha}{\|z^\alpha\|}\,, \tag{2.5}$$

where the coefficients λ_α are again independent complex Gaussian random variables with mean zero and variance one.

Our simplest result concerns the the expected density $\mathbf{E}_{|P}(Z_{f_1,\ldots,f_m})$ of the simultaneous zeros of $(f_1, \ldots, f_m)$ chosen independently from $H^0(\mathbb{CP}^m, \mathcal{O}(p), P)$. It is the measure on $\mathbb{C}^{*m}$ given by

$$\begin{aligned} &\mathbf{E}_{|P}(Z_{f_1,\ldots,f_m})(U) \\ = &\int d\gamma_{p|P}(f_1) \cdots \int d\gamma_{p|P}(f_m) \left[\#\{z \in U : f_1(z) = \cdots = f_m(z) = 0\}\right], \end{aligned} \tag{2.6}$$

for $U \subset \mathbb{C}^{*m}$, where the integrals are over $H^0(\mathbb{CP}^m, \mathcal{O}(p), P)$. We will determine the asymptotics of the expected density as the polytope is dilated $P \to NP, N \in \mathbb{N}$.

Theorem 2 [19] *Suppose that P is a simple polytope in $\mathbb{R}^m$. Then, as P is dilated to NP,*

$$\frac{1}{(N\delta)^m}\mathbf{E}_{|NP}(Z_{f_1,\dots,f_m}) \to \begin{cases} \omega_{\mathrm{FS}}^m & \text{on } \mathcal{A}_P \\ 0 & \text{on } \mathbb{C}^{*m} \setminus \mathcal{A}_P \end{cases},$$

in the distribution sense; i.e., for any open $U \subset \mathbb{C}^{*m}$, we have

$$\frac{1}{(N\delta)^m}\mathbf{E}_{|NP}\big(\#\{z \in U : f_1(z) = \cdots = f_m(z) = 0\}\big) \to m!\mathrm{Vol}_{\mathbb{CP}^m}(U \cap \mathcal{A}_P)\,.$$

There are also results for $k < m$ polynomials. The distribution of zeros is ω_{FS}^k in $\mathcal{A}_P$ as if there were no constraint, while there is an exotic distribution in $\mathbb{C}^{*m} \setminus \mathcal{A}_P$ which depends on the exponentially decaying asymptotics of the conditional Bergman- Szegö kernel

$$\Pi_{|NP}(z,w) = \sum_{\alpha \in NP} \frac{z^\alpha \overline{w}^\alpha}{||z^\alpha||_{FS}||w^\alpha||_{FS}},$$

i.e. the orthogonal projection onto the subspace (2.4). It is obtained by sifting out terms in the (elementary) Szegö projector of $H^0(\mathbb{CP}^m, \mathcal{O}(pN))$ using the polytope character $\chi_{NP}(e^{i\varphi}) = \sum_{\alpha \in NP} e^{i\langle \alpha, \varphi\rangle}$. To obtain asymptotics in the forbidden region, we write $\chi_{NP}(e^{i\varphi}) = \int_{M_P} \Pi_N^{M_P}(e^{i\varphi}w, w)dV(w)$, where Π^{M_P} is Bergman-Szegö kernel of the toric variety M_P associated to P. We then make an explicit construction of $\Pi_N^{M_P}$ as a complex oscillatory integral. An alternative is to express χ_{NP} as a Todd derivative of an exponential integral over P (following works of Khovanskii-Pukhlikov, Brion-Vergne and Guillemin). We thus obtain a complex oscillatory integral formula for $\Pi_{|NP}(z,w)$. To obtain asymptotics in the forbidden region we carefully deform the contour into the complex and apply a complex stationary phase method.

Although we only discuss expected behavior of zeros here, the distribution of zeros is *self-averaging*: i.e., almost all polynomials exhibit the expected behavior in an asymptotic sense. We also expect similar results for critical points.

3. Eigenfunctions

We now turn to the eigenvalue problem $\Delta_g \varphi_\nu = \lambda_\nu^2 \varphi_\nu$, $\langle \varphi_\nu, \varphi_{\nu'}\rangle = \delta_{\nu\nu'}$ on a compact Riemannian manifold (M, g). We denote the λ-eigenspace by V_λ. The role of the Szegö kernel is now played by the kernel $E_\lambda(x,y) = \sum_{\lambda_\nu \leq \lambda} \varphi_\nu(x)\overline{\varphi_\nu(y)}$ of the spectral projections.

3.1. L^p bounds

Our first concern is with L^p norms of L^2-normalized eigenfunctions. We measure the growth rate of L^p norms by $L^p(\lambda, g) = \sup_{\varphi \in V_\lambda : ||\varphi||_{L^2}=1} ||\varphi||_{L^p}$. By the local Weyl law, $E_\lambda(x,x) = \sum_{\lambda_\nu \leq \lambda} |\varphi_\nu(x)|^2 = (2\pi)^{-n} \int_{|\xi| \leq \lambda} d\xi + O(\lambda^{n-1})$, it follows that $L^\infty(\lambda, g) = 0(\lambda^{\frac{n-1}{2}})$ on any compact Riemannian manifold. This bound, which is based entirely on a local analysis, is sharp in the case of the standard round sphere, S^n or on any rotationally invariant metric on S^2, but is far off in the case of flat tori. This motivates:

- **Problem 5** For which (M, g) is this estimate sharp? Which (M, g) are extremal for growth rates of $||\varphi_\lambda||_p$, both maximal and minimal? What if M has a boundary? What is the expected L^p norm of a 'random' L^2-normalized polynomial or eigenfunction?

In [20], we give a necessary condition for *maximal* eigenfunction growth: there must exist a point $x \in M$ for which the set $\mathcal{L}_x = \{\xi \in S_x^* M : \exists T : \exp_x T\xi = x\}$ of directions of geodesic loops at x has positive surface measure.

Theorem 3 [20] *If $\mathcal{L}_x$ has measure 0 in $S_x^* M$ for every $x \in M$ then*

$$L^p(\lambda, g) = o(\lambda^{\delta(p)}), \quad p > \tfrac{2(n+1)}{n-1} \quad \delta(p) = \begin{cases} n(\frac{1}{2} - \frac{1}{p}) - \frac{1}{2}, & \frac{2(n+1)}{n-1} \leq p \leq \infty \\ \frac{n-1}{2}(\frac{1}{2} - \frac{1}{p}), & 2 \leq p \leq \frac{2(n+1)}{n-1}. \end{cases} \tag{3.1}$$

The L^p-bounds $O(\lambda^{\delta(p)})$ were proved by Sogge to hold for all (M, g).
We further prove:

Theorem 4 [20] *(see also [17]) Suppose that (M, g) is:*

- *Real analytic and that $L^\infty(\lambda, g) = \Omega(\lambda^{(n-1)/2})$. Then (M, g) is a Y_ℓ^m - manifold, i.e. $\exists m$ such that all geodesics issuing from the point m return to m at time ℓ. In particular, if dim $M = 2$, then M is topologically a 2-sphere S^2 or a real projective plane $\mathbb{R}P^2$.*
- *Generic. Then $L^\infty(\lambda, g) = o(\lambda^{(n-1)/2})$.*

Here, $\Omega(\lambda^{\frac{n-1}{2}})$ means $O(\lambda^{\frac{n-1}{2}})$ but not $o(\lambda^{\frac{n-1}{2}})$. The generic result holds because $\mathcal{L}_x$ has measure 0 in $S_x^* M$ for all $x \in M$ for a residual set of metrics.

In the case of random polynomials, or random combinations of eigenfunctions in short spectral intervals, the almost sure growth of L^∞ norms is $O(\sqrt{\log N})$ while the L^p norms for $p < \infty$ are bounded. This was proved by J. Vanderkam [24] for S^m, Nonnenmacher-Voros [14] for elliptic curves and Shiffman-Zelditch (to appear) for the general case using Levy concentration of measure estimates.

3.2. Integrable case

Results on *minimal growth* have been obtained by J. A. Toth and the author in the *quantum completely integrable* case, where $\sqrt{\Delta} = P_1$ commutes with $n-1$ first order pseudodifferential operators $P_2, \ldots, P_n \in \Psi^1(M)$ ($n = \dim M$) satisfying

$[P_i, P_j] = 0$ and whose symbols define a moment map $\mathcal{P} := (p_1, \dots, p_n)$ satisfying $dp_1 \wedge dp_2 \wedge \cdots \wedge dp_n \neq 0$ on a dense open set $\Omega \subset T^*M - 0$. Since $\{p_i, p_j\} = 0$, the functions $p_1, \dots, p_n$ generate a homogeneous Hamiltonian $\mathbb{R}^n$-action whose orbits foliate $T^*M - 0$. We refer to this foliation as the *Liouville foliation.*

We consider the L^p norms of the L^2-normalized joint eigenfunctions $P_j\varphi_\lambda = \lambda_j\varphi_\lambda$. The spectrum of Δ often has bounded multiplicity, so the behaviour of joint eigenfunctions has implications for all eigenfunctions.

Theorem 5 [22, 23] *Suppose that the Laplacian Δ_g of (M, g) is quantum completely integrable and that the joint eigenfunctions have uniformly bounded L^∞ norms. Then (M, g) is a flat torus.*

This is a kind of quantum analogue of the 'Hopf conjecture' (proved by Burago-Ivanov) that metrics on tori without conjugate points are flat. In [23], a quantitative improvement is given under a further non-degeneracy assumption. Unless (M, g) is a flat torus, the Liouville foliation must possess a singular leaf of dimension $< n$. Let ℓ denote the minimum dimension of the leaves. We then construct a sequence of eigenfunctions satisfying:

$$\|\varphi_k\|_{L^\infty} \geq C(\epsilon)\lambda_k^{\frac{n-\ell}{4}-\epsilon}, \quad \|\varphi_k\|_{L^p} \geq C(\epsilon)\lambda_k^{\frac{(n-\ell)(p-2)}{4p}-\epsilon}, \quad (2 < p)$$

for any $\epsilon > 0$. It is easy to construct examples were $\ell = n-1$, but it seems plausible that in 'many' cases $\ell = 1$. To investigate this, one would study the boundary faces of the image $\mathcal{P}(T^*M - 0)$ of $T * M - 0$ under a homogeneous moment map. For a related study in the case of torus actions, see Lerman-Shirokova [12].

3.3. Quantum ergodicity

Quantum ergodicity is concerned with the sums ($A \in \Psi^0(M)$):

$$S_p(\lambda) = \sum_{\nu:\lambda_\nu \leq \lambda} |\langle A\varphi_\nu, \varphi_\nu\rangle - \omega(A)|^p, \quad \omega(A) = \frac{1}{Vol(S^*M)} \int_{S^*M} \sigma_A d\mu. \tag{3.2}$$

In work of A.I. Schnirelman [11], Colin de Verdiere and the author [27], it is shown that $S_p(\lambda) = o(N(\lambda))$ if G^t is ergodic. In the author's view [27], this is best viewed as a convexity theorem. We mention briefly some new results.

In work of Gerard-Leichtnam [7] and Zelditch-Zworski [30], the ergodicity result was extended to domains $\partial\Omega$ with piecewise smooth boundary and ergodic billiard flow. Since the billiard map on $B^*\partial\Omega$ is ergodic whenever the billiard flow is, suitable boundary values of ergodic eigenfunctions (e.g. $\varphi_k|_{\partial\Omega}$ in the Neumann case or $\partial_\nu\varphi_k|_{\partial\Omega}$ in the Dirichlet case) should also have the ergodic property. This was conjectured by S. Ozawa in 1993. A proof is given in our work with A. Hassell [8] for convex piecewise smooth domains with ergodic billiards (in the case of domains with Lipschitz normal and with Dirichlet boundary conditions, this had earlier been proved in [7] by a different method).

Little is known about the rate of decay. For sufficiently chaotic systems (satisfying the central limit theorem) one can get the tiny improvement $S_p(\lambda) =$

$O(N(\lambda)/(\log\lambda)^p))$ [28]. The asymptotics $S_2(\lambda) \sim B(A)\lambda$ have recently been obtained by Luo-Sarnak [13] for Hecke eigenfunctions of the modular group, exploiting the connections with L-functions. These asymptotics (though not the coefficient) are predicted by the random polynomial model. Other strong bounds in the arithmetic case were obtained by Kurlberg-Rudnick for eigenfunctions of certain quantized torus automorphisms [10]. Bourgain-Lindenstrauss [3] and Wolpert [25] have developed the 'non-scarring' result of [16] to give entropy estimates of possible quantum limit measures in arithmetic cases.

A natural problem is the converse:

- **Problem 6** What can be said of the dynamics if $S_p(\lambda) = o(N(\lambda))$? Does quantum ergodicity imply classical ergodicity?

It is known that classical ergodicity is equivalent to this bound plus estimates on off-diagonal terms [21]. The existence of KAM quasimodes (due to Lazutkin [11], Colin de Verdiere [5], and Popov [15]) makes it very likely that KAM systems are *not* quantum ergodic, nor are (M, g) which have stable elliptic orbits.

A further problem which may be accessible is:

- **Problem 7** How are the nodal sets $\{\varphi_\nu = 0\}$ distributed in the limit $\nu \to \infty$?

In [14] (for elliptic curves) and [19] (general Kähler manifolds) it is proved that the *complex zeros* of quantum ergodic eigenfunctions become uniformly distributed relative to the volume form. Can one prove an analogue for the real zeros?

References

[1] P. Bleher, B. Shiffman and S. Zelditch, Universality and scaling of correlations between zeros on complex manifolds, *Invent. Math.* 142 (2000), 351–395.

[2] P. Bleher, B. Shiffman and S. Zelditch, Correlations between zeros and supersymmetry, Commun. Math. Phys. 224 (2001) 1, 255-269.

[3] J. Bourgain and E. Lindenstrauss, Entropy of quantum limits (preprint, 2002).

[4] Boutet de Monvel, L.; Sjstrand, J. Sur la singularit des noyaux de Bergman et de Szeg.Journes: quations aux Drives Partielles de Rennes (1975), 123–164. Asterisque, No. 34-35, Soc. Math. France, Paris, 1976.

[5] Colin de Verdire, Yves Quasi-modes sur les varits Riemanniennes. Invent. Math. 43 (1977), no. 1, 15–52.

[6] H. Donnelly, C. Fefferman, Nodal sets of eigenfunctions on Riemannian manifolds. Invent. Math. 93 (1988), no. 1, 161–183.

[7] P. Gerard and E. Leichtnam, Ergodic properties of eigenfunctions for the Dirichlet problem, Duke Math. J. 71 (1993), 559–607.

[8] A. Hassell and S. Zelditch, Quantum ergodicity and boundary values of eigenfunctions (preprint, 2002).

[9] D. Jakobson, N. Nadirashvili, and J. Toth, Geometry of eigenfunctions (to appear in Russian Math Surveys).

[10] P. Kurlberg and Z. Rudnick, Value distribution for eigenfunctions of desymmetrized quantum maps. Internat. Math. Res. Notices 2001, no. 18, 985–1002.

[11] V. F. Lazutkin, *KAM theory and semiclassical approximations to eigenfunctions. With an addendum by A. I. Shnirelman.* Ergebnisse der Mathematik und ihrer Grenzgebiete 24. Springer-Verlag, Berlin, 1993.
[12] E. Lerman and N. Shirokova, Completely integrable torus actions on symplectic cones, Math. Res. Letters 9 (2002), 105-116.
[13] W. Luo and P. Sarnak, Ergodicity of eigenfunctions on $SL(2,Z) \setminus H$, II (in preparation).
[14] Nonnenmacher, S.; Voros, A. Chaotic eigenfunctions in phase space. J. Statist. Phys. 92 (1998), no. 3-4, 431–518.
[15] G. Popov, Invariant tori, effective stability, and quasimodes with exponentially small error terms. I. Birkhoff normal forms, Ann. Henri Poincare 1 (2000), 223–248.
[16] Z. Rudnick and P. Sarnak, The behaviour of eigenstates of arithmetic hyperbolic manifolds. Comm. Math. Phys. 161 (1994), no. 1, 195–213.
[17] Yu. G. Safarov, Asymptotics of a spectral function of a positive elliptic operator without a nontrapping condition, Funct. Anal. Appl. 22 (1988), no. 3, 213–223.
[18] B. Shiffman and S. Zelditch, Distribution of zeros of random and quantum chaotic sections of positive line bundles. Comm. Math. Phys. 200 (1999), no. 3, 661–683.
[19] B. Shiffman and S. Zelditch, Random polynomials with prescribed Newton polytope I (preprint, 2002).
[20] C. Sogge and S. Zelditch, Riemannian manifolds with maximal eigenfunction growth (Duke Math J.).
[21] T. Sunada, Quantum ergodicity. Progress in inverse spectral geometry, 175–196, Trends Math., Birkhuser, Basel, 1997.
[22] J. A. Toth and S. Zelditch, Riemannian manifolds with uniformly bounded eigenfunctions, Duke Math. J. 111 (2002), 97-132.
[23] J. A. Toth and S. Zelditch, L^p estimates of eigenfunctions in the completely integrable case (preprint, 2002).
[24] J. M. VanderKam, L^∞ norms and quantum ergodicity on the sphere. IMRN 7 (1997), 329–347.
[25] S. A. Wolpert, The modulus of continuity for $\Gamma_0(m)\backslash\mathbb{H}$ semi-classical limits. Comm. Math. Phys. 216 (2001), no. 2, 313–323.
[26] S. Zelditch, From random polynomials to symplectic geometry, in *XIIIth International Congress of Mathematical Physics,* International Press (2001), 367-376.
[27] S. Zelditch, Quantum ergodicity of C^* dynamical systems. Comm. Math. Phys. 177 (1996), no. 2, 507–528.
[28] S. Zelditch, On the rate of quantum ergodicity. I. Upper bounds. Comm. Math. Phys. 160 (1994), 81–92.
[29] S. Zelditch, Szegö kernels and a theorem of Tian. IMRN 6 (1998), 317–331.
[30] S. Zelditch and M. Zworski, Ergodicity of eigenfunctions for ergodic billiards. Comm. Math. Phys. 175 (1996), no. 3, 673–682.

ICM 2002 · Vol. II · 743–753

Some Results Related to Group Actions in Several Complex Variables

Xiangyu Zhou*

Abstract

In this talk, we'll present some recent results related to group actions in several complex variables. We'll not aim at giving a complete survey about the topic but giving some our own results and related ones.

We'll divide the results into two cases: compact and noncompact transformation groups. We emphasize some essential differences between the two cases. In the compact case, we'll mention some results about schlichtness of envelopes of holomorphy and compactness of automorphism groups of some invariant domains. In the noncompact case, we'll present our solution of the longstanding problem – the so-called extended future tube conjecture which asserts that the extended future tube is a domain of holomorphy. Invariant version of Cartan's lemma about extension of holomorphic functions from the subvarities in the sense of group actions will be also mentioned.

2000 Mathematics Subject Classification: 32.
Key words and phrases: Domain of holomorphy, Plurisubharmonic function, Group actions.

1. Fundamentals of several complex variables

About one century ago, Hartogs discovered that there exist some domains in several complex variables on which any holomorphic functions can be extended to larger domains, being different with one complex variable. This causes a basic concept – domain of holomorphy.

Definition. *A domain of holomorphy in $\mathbb{C}^n$ is a domain on which there exists a holomorphic function which can't be extended holomorphically across any boundary points.*

A domain in $\mathbb{C}^n$ is called holomorphically convex, if given any infinite discrete point sequence z_k there exists a holomorphic function f s.t. $f(z_k)$ is unbounded (or $|f(x_v)| \to +\infty$). Consequently, there exists a holomorphic function which tends to

*Institute of Mathematics, AMSS, Chinese Academy of Sciences, Beijing; Department of Mathematics, Zhejiang University, Hangzhou, China. E-mail: xyzhou@math08.math.ac.cn

$+\infty$ at the boundary. By Cartan-Thullen's theorem, a domain in $\mathbb{C}^n$ is a domain of holomorphy if and only if the domain is Stein, i.e., holomorphically convex.

Definition. *A function φ with value in $[-\infty, +\infty)$ on the domain D in $\mathbb{C}^n$ is called plurisubharmonic (p.s.h.): if (i) φ is upper semicontinuous (i.e., $\{\varphi < c\}$ is open for each $c \in \mathbb{R}$, or equivalently $\overline{\lim}_{z \to z_0} \varphi(z) \leq \varphi(z_0)$ for $z_0 \in D$); (ii) for each complex line $L := \{z_0 + tr : z_0 \in D\}, \varphi|_{L \cap D}$ is subharmonic w.r.t. one complex variable t.*

An equivalent definition in the sense of distributions is that $i\partial\overline{\partial}\varphi$ is a positive current; in particular, when φ is C^2, this means Levi form $\left(\frac{\partial^2 \varphi}{\partial z_i \partial \overline{z}_j}\right) \geq 0$ everywhere. In other words, $dJd\varphi \geq 0$, where J is the complex structure. (If $i\partial\overline{\partial}\varphi > 0$, then φ is called strictly p.s.h.)

Example. For a bounded domain or a domain biholomorphic to a bounded domain, the Bergman kernel $K(z, \overline{z})$ is strictly p.s.h..

A pseudoconvex domain in $\mathbb{C}^n$ is a domain on which there exists a p.s.h. function which tends to $+\infty$ at the boundary. It's easy to see that a holomorphical convex domain is pseudoconvex, since $|f|^2$ is plurisubharmonic function where f is given in the consequence of the definition of a Stein domain.

A deep characterization of a domain of holomorphy is given by a solution to Levi problem which is the converse of the above statement.

Fact. A domain D in $\mathbb{C}^n$ is a domain of holomorphy if and only if the domain is pseudoconvex.

A natural corresponding concept of the domain of holomorphy in the setting of complex manifolds (complex spaces) is the so-called Stein manifold (Stein space), which is defined as a holomorphically convex and holomorphically separable complex manifold (space) . A complex manifold (or space with finite embedding dimension) is Stein if and only if it is a closed complex submanifold (or subvariety) in some $\mathbb{C}^n$, and if and only if there exists a strictly p.s.h. exhaustion function which is $\mathbb{R}$-valued (i.e., the value $-\infty$ is not allowed). A complex reductive Lie group, in particular a complex semisimple Lie group, is a Stein manifold.

We know that a domain of holomorphy or a Stein manifold are defined by special holomorphic functions which are usually hard to construct in several complex variables. However, a pseudoconvex domain is defined by a special p.s.h. function which is a real function and then relatively easy to construct. Construction of various holomorphic objects in several complex variables and complex geometry is a fundamental and difficult problem. An important philosophy here is reducing the construction of holomorphic functions to the construction of plurisubharmonic functions, because of the solution of Levi problem and Hörmander's L^2 estimates for $\bar{\partial}$ and other results.

2. Group actions in several complex variables

Definition. *A group action of the group G on a set X is given by a mapping $\varphi : G \times X \to X$ satisfying the following: 1) $e \cdot x = x$, 2) $(ab) \cdot x) = a \cdot (b \cdot x)$, where e is the identity of the group, $a, b, \in G, x \in X$, $a \cdot x := \varphi(a, x)$.*

A group action on a set can be restricted on various cases. When the set is a topological space and the group is a topological group, the action is continuous, then one gets a topological transformation group; when the space is a metric space, the transformation preserves the metric, then one gets a motion group; when the set is a differentiable manifold and the group is a Lie group, the action is differentiable, then one gets a Lie transformation group; when the set is a vector space, the transformation preserves the vector space structure, then one gets a linear transformation group; when the set is an algebraic variety (or a scheme), the group is an algebraic group, and the action is algebraic, one gets an algebraic transformation group; when the set is a complex space, the transformation is holomorphic, and the action is real analytic, then one gets a (real) holomorphic transformation group (note that in this case, if the action is continuous then it is also real analytic); if the set is a complex space, the group is a complex Lie group, and the the action is holomorphic, then one gets a complex (holomorphic) transformation group.

In this talk, we're mainly concerned with the last case. We consider a complex Lie group $G_{\mathbb{C}}$ with a real form $G_{\mathbb{R}}$ acting holomorphically on a complex manifold (also called holomorphic $G_{\mathbb{C}}$- manifold) and a $G_{\mathbb{R}}$-invariant domain. It's known that a complex reductive Lie group has a unique maximal compact subgroup up to conjugate as its real form, but it also has many noncompact real forms.

A group action on a set can be regarded as a representation of the group on the whole group of transformations. An effective group action means the representation is faithful, so it corresponds to a (closed) subgroup of the whole transformation group.

Actually, many domains in several complex variables such as Hartogs, circular, Reinhardt and tube domains can be formulated in the setting of group actions.

Examples. a) Hartogs and circular domains: consider the Hartogs action of $\mathbb{C}^*$ with the real form S^1 on $\mathbb{C}^n$: $\mathbb{C}^* \times \mathbb{C}^n \to \mathbb{C}^n$ given by $(t,(z_1,\cdots,z_n)) \to (tz_1,z_2,\cdots,z_n)$, then Hartogs domain is S^1-invariant domain; consider the circular action of $\mathbb{C}^*$ with the real form S^1 on $\mathbb{C}^n$: $\mathbb{C}^* \times \mathbb{C}^n \to \mathbb{C}^n$ given by $(t,(z_1,\cdots,z_n)) \to (tz_1,tz_2,\cdots,tz_n)$, then circular domain is S^1-invariant domain.

b) Reinhardt domains: consider the Reinhardt action of $(\mathbb{C}^*)^n$ on $\mathbb{C}^n$ given by

$$((t_1,\cdots,t_n),(z_1,\cdots,z_n)) \to (t_1z_1,\cdots,t_nz_n),$$

then Reinhardt domain is $(S^1)^n$-invariant domain. One can similarly defines matrix Reinhardt domains

c) tube domains: consider the action of $\mathbb{R}^n$ on $\mathbb{C}^n$ given by $(r,z) \to r+z$, then $\mathbb{R}^n$-invariant domain is tube domain.

d) future tube: let M^4 be the Minkowski space with the Lorentz metric: $x \cdot y = x_0y_0 - x_1y_1 - x_2y_2 - x_3y_3$, where $x = (x_0,x_1,x_2,x_3), y = (y_0,y_1,y_2,y_3) \in R^4$; let V^+ and $V^- = -V^+$ be the future and past light cones in R^4 respectively, i.e. $V^\pm = \{y \in M : y^2 > 0, \pm y_0 > 0\}$, the corresponding tube domains $\tau^\pm = T^{V^\pm} = R^4 + iV^\pm$ in $\mathbb{C}^4$ are called future and past tubes; let L be the Lorentz group, i.e. $L = O(1,3)$, L has four connected components, denote the identity component of L by $L_+^\uparrow$, which is called the restricted Lorentz group, i.e. $L_+^\uparrow = SO_+(1,3)$; let $L(\mathbb{C})$ be the complex Lorentz group, i.e.$L = O(1,3,\mathbb{C}) \cong O(4,\mathbb{C}), L(\mathbb{C})$ has two

connected components, denote the identity component of $L(\mathbb{C})$ by $L_+(\mathbb{C})$, called the proper complex Lorentz group which has the restricted Lorentz group as its real form. Considering the linear action of $L_+(\mathbb{C})$ on $\mathbb{C}^4$, the future (or past) tube is $L_+^{\uparrow}$-invariant.

Denote the N-point future tube by $\tau_N^{\pm} = \tau^{\pm} \times \cdots \times \tau^{\pm}$ N-times, let $L_+(\mathbb{C})$ act diagonally on $\mathbb{C}^{4N}$, i.e. for $z = (z^{(1)}, \cdots, z^{(N)}) \in \mathbb{C}^{4N}, \wedge z = (\wedge z^{(1)}, \cdots, \wedge z^{(N)})$ where $\wedge \in L_+(\mathbb{C})$, then $\tau_N^{\pm}$ is $L_+^{\uparrow}$,-invariant.

e) matrix Reinhardt domains: let $\mathbb{C}^n[m \times m] = \{(Z_1, \cdots, Z_n)$: $Z_j \in \mathbb{C}[m \times m]\}$ be the space of n-tuples of $m \times m$ matrices. A domain $D \subset C^n[m \times m]$ is called matrix Reinhardt if it is invariant under the diagonal $\mathrm{U}(m) \times \mathrm{U}(m)$ action $(U, V)(Z_1, \cdots, Z_n) \mapsto (UZ_1V, \cdots, UZ_nV)$. These domains are the usual Reinhardt domains in the case $m = 1$. $\mathrm{Diag}(D)$ is defined as the intersection of D with the diagonal matrices $(Z_1, \cdots, Z_n) \in \mathbb{C}^n[m \times m]$

Slice theory

When G is a Lie transformation group properly acting on a smooth manifold X (e.g. when G is compact), one has a satisfactory slice theory about the structure of a neighborhood of an orbit. This theory was extended to the case of an affine reductive group action regularly on an affine variety by D. Luna ([20]) and the case of a complex reductive Lie group G action holomorphically on a Stein space X by Snow ([27]). In these cases, the structure of a neighborhood of a closed orbit is finely determined. We state the result for reduced Stein spaces. Let $G \cdot x$ be a closed orbit, then there exists a locally closed G_x-invariant Stein subspace B containing x s.t. the natural map from the homogeneous fiber bundle $G \times_{G_x} B$ over $G/G_x \cong G \cdot x$ is biholomorphic onto a π-saturated open Stein subset of X, where $\pi : X \to X//G$ is the categorical quotient (or GIT quotient) which exists as a Stein space. Here B is called a slice at x. The slice B is transversal to the closed orbit $G \cdot x$. When X is regular at x, then B can be chosen to be regular.

As a consequence of the slice theorem, one has a stratification of the categorical quotient $X//G$ at least when X is a Stein manifold. The stratum with maximal dimension is Zariski open in $X//G$ and is contained in the regular part of $X//G$. This is called principal stratum. The inverse of the principal stratum under $\pi : X \to X//G$ consists of all G-closed orbits satisfying that they are of maximal dimension k among the dimensions of all G-closed orbits and their corresponding isotropy groups are of minimum number of components. Such orbits are called principal closed orbits, and the corresponding isotropy groups are called principal. When $k = \dim G$, then X is called having FPIG.

3. Some results on compact holomorphic transformation groups

The relationship between orbit connectedness, orbit convexity, and holomorphical convexity goes back to the beginning of this century, when several complex variables was born. Due to Hartogs, Reinhardt, H.Cartan and others, one already knew some classical relations between completeness, logarithmic convexity and holo-

morphical convexity for circular domains, Hartogs domains, and Reinhardt domains. The orbit connectedness and orbit convexity are defined in a general setting (for arbitrary compact connected Lie group), which correspond to completeness and logarithmic convexity when one restricts to the above domains.

There are some fundamental relationships between orbit connectedness and orbit convexity with holomorphically convexity and envelope of holomorphy for invariant domains.

Definition. *Let $G_{\mathbb{C}}$ be a connected complex Lie group, $G_{\mathbb{R}}$ be a connected closed real form of $G_{\mathbb{C}}$. Let X be a holomorphic $G_{\mathbb{C}}$-space, $D \subset X$ be a $G_{\mathbb{R}}$-invariant set, we call D orbit connected, if for $b_z : G_{\mathbb{C}} \to X, g \mapsto g \cdot z, b_z^{-1}(D)$ is connected for each $z \in D$. When $(G_{\mathbb{C}}, G_{\mathbb{R}})$ is a geodesic convex pair(i.e. the map* $\mathrm{Lie}(\mathrm{G}_{\mathbb{R}}) \times \mathrm{G}_{\mathbb{R}} \ni (\mathrm{v}, \mathrm{g}) \to \exp(\mathrm{iv})\mathrm{g} \in \mathrm{G}_{\mathbb{C}}$ *is a homeomorphism, cf. [3]), D is called orbit convex if for each $z \in D$, and $v \in i\mathrm{Lie}(\mathrm{G}_{\mathbb{R}})$ s.t. $\exp(v) \in b_z^{-1}(D)$ it follows $\exp(tv) \in b_z^{-1}(D)$ for all $t \in [0, 1]$.*

Roughly speaking, orbit connectedness means that $G_{\mathbb{C}}x \cap D$ is connected for every $x \in D$.

One has known for a long time that the envelope of holomorphy of a domain in $\mathbb{C}^n$ (or more general a Riemann domain over $\mathbb{C}^n$) exists uniquely as a Riemann domain over $\mathbb{C}^n$. There is a difficult problem of univalence: When is the envelope of holomorphy of a domain in $\mathbb{C}^n$ itself a domain in $\mathbb{C}^n$? We have the following criteria for the univalence of the envelope of holomorphy for certain invariant domains:

Theorem 1 ([36]). *Let X be a Stein manifold, $K^{\mathbb{C}}$ be a complex reductive Lie group holomorphically acting on X, where K is a connected compact Lie group and $K^{\mathbb{C}}$ be its universal complexification. Let $D \subset X$ be a K-invariant orbit connected domain. Then the envelope of holomorphy $E(D)$ of D is schlicht and orbit convex if and only if the envelope of holomorphy $E(K^{\mathbb{C}} \cdot D)$ of $K^{\mathbb{C}} \cdot D$ is schlicht. Furthermore, in this case, $E(K^{\mathbb{C}} \cdot D) = K^{\mathbb{C}} \cdot E(D)$.*

When $K = S^1$ and the action is circular (or α-circular) and Hartogs, the corresponding concepts of orbit connectedness for such domains were introduced separately and the above results were obtained and stated separately by Casadio Tarabushi and Trapani in [1,2].

When $K = (S^1)^n$ and the action is Reinhardt, the result is well known as a classical result about Reinhardt domain which asserts that any Reinhardt domain in $(\mathbb{C}^*)^n$ has schlicht envelope of holomorphy.

Some other results were also included in the above theorem. So our result can also be regarded as an extension of their results and the classical result on Reinhardt domains in a unified way.

In the proof, a theorem due to Harish-Chandra on the infinite dimensional representation of Lie groups plays an important role.

We also give some examples of orbit connected domains. Let $X = K^{\mathbb{C}}/L^{\mathbb{C}}$, the action of $K^{\mathbb{C}}$ on X be given by the left translations. When L is connected or (K, L) is a symmetric pair, then any K-invariant domain is orbit connected. The simplest example is Reinhardt domains in $(\mathbb{C}^*)^n$.

The origin of orbit connectedness could at least go back to [28].

Example. A theorem of V.Bargmann, D. Hall and A.S. Wightman (cf.

Wightman [32], Jost [12], Streater-Wightman [28]) asserts that τ_N^+ is orbit connected.

We also consider the homogeneous embeddings of $K^{\mathbb{C}}/L^{\mathbb{C}}$. Let X be a smooth homogeneous space embedding of $K^{\mathbb{C}}/L^{\mathbb{C}}$, $D \subset X$ be a K-domain. Assume that L is connected or (K, L) is a symmetric pair. Then $E(D)$ is schlicht and orbit convex. In particular, every matrix Reinhardt domain of holomorphy D is orbit convex. Since an orbit convex matrix Reinhardt domain has a path connected $\mathrm{Diag}(D)$, so a matrix Reinhardt domain of holomorphy has a connected $\mathrm{Diag}(D)$.

Theorem 2([37]). *Let K be a connected compact Lie group, L be a closed (not necessarily connected) subgroup of K. Let $K^{\mathbb{C}}$ and $L^{\mathbb{C}}$ be respectively universal complexification of K and L. Suppose that D is K-invariant relatively compact domain in $K^{\mathbb{C}}/L^{\mathbb{C}}$ (Here the action of $K^{\mathbb{C}}$ is given by left translations). Then (i) $Aut(D)$ is a compact Lie group; (ii) Any proper holomorphic self-mapping of D is biholomorphic if K is semisimple or a direct product of a semisimple compact Lie group and a compact torus.*

By a result of Matsushima, $K^{\mathbb{C}}/L^{\mathbb{C}}$ is a Stein manifold which is a holomorphic $K^{\mathbb{C}}$ - manifold w.r.t. left translation action.

The motivations of the present work are two-folds: the result (i) is to extend a main result of [4], where the same result was obtained by requiring a restrictive condition that (K, L) is a symmetric pair,i.e., K/L is a compact Riemannian symmetric space; the result (ii) is to extend a classical result which asserts that proper self mapping of the relatively compact Reinhardt domains in $(\mathbb{C}^*)^n$ is biholomorphic.

The proof is involved with many famous results such as Mostow decomposition theorem, H. Cartan's theorem about compactness of automorphism groups, Andreotti-Frankel's theorem on homology group of a Stein manifold, the holomorphic version of de Rham's theorem on a Stein manifold, a result of Milnor's about CW complex, a result from iteration theory, Poincaré duality theorem, degree theory for proper mappings, covering lifting existence theorem, and a result about compact semisimple Lie groups et al.

4. Extended future tube conjecture

Let's keep the notation in Example d of the section 2. The set $\tau'_N := \{\wedge z : z \in \tau_N^+, \wedge \in L_+(\mathbb{C})\}$ is called the extended future tube.

The extended future tube conjecture, which arose naturally from axiomatic quantum field theory at the end of 1950's, asserts that the extended future tube τ'_N is a domain of holomorphy for $N \geq 3$. This conjecture turns out to be very beautiful and natural. In their papers, Vladimirov and Sergeev said that the importance of the conjecture is also due to the fact that there are some assertions in QFT, such as the finite covariance theorem of Bogoliubov-Vladimirov, proved only assuming that this conjecture is true.

According to the axiomatic quantum field theory (cf. [12,13,28]), one may describe physical properties of a quantum system using the Wightman functions which correspond to holomorphic functions in τ_N^+ invariant w.r.t. the diagonal action of $L_+^{\uparrow}$. This sort of functions have the following extension property.

BHW Theorem (due to Bargman, Hall, and Wightman 1957). An $L_+^{\uparrow}$-invariant holomorphic function on τ_N^+ can be extended to an $L_+(\mathbb{C})$-invariant holomorphic function on τ_N' (cf. [12,13,28]).

In the proof, the orbit connectedness of τ_N^+ play a key role. With this and Identity Theorem, one can easily define the invariant holomorphic extension.

So, a natural question arises, i.e., can these holomorphic functions be extended further? Or, is τ_N' holomorphic convex w.r.t. $L_+(\mathbb{C})$-invariant holomorphic function? After some argument, this is equivalent to ask if τ_N' is a domain of holomorphy.

Streater's theorem. A holomorphic function on the Dyson domain $\tau_N^+ \cup \tau_N^- \cup J$ (where $J := \tau_N' \cap M^{4N}$ is the set of Jost points which was proved to exist and characterized by R. Jost) can be extended to a holomorphic function on τ_N' (cf. [12,28]).

So, a natural question is to construct the envelope of holomorphy of the Dyson domain $\tau_N^+ \cup \tau_N^- \cup J$ (This question is mentioned in the article "Quantum field theory" of the Russian's great dictionary "Encyclopedia of Mathematics"). That the extended future tube conjecture holds is equivalent to that this envelope of holomorphy is exactly the extended future tube τ_N'.

The conjecture have been mentioned as an open problem in many classical ([12,28]) and recent references ([11,21-24,28-31]) and references therein. In [38,39], we found a route to solve the conjecture via Kiselman-Loeb's minimum principle and Luna's slice theory. Let's recall the minimum principle.

Minimum principle

Let X be a complex manifold, $G_{\mathbb{C}}$ a connected complex Lie group, $G_{\mathbb{R}}$ a connected closed real form of $G_{\mathbb{C}}$. Denote $\psi : G_{\mathbb{C}} \to G_{\mathbb{C}}/G_{\mathbb{R}}$, and $p : X \times G_{\mathbb{C}} \to X$ the natural projections.

$G_{\mathbb{C}}$ acts on $X \times G_{\mathbb{C}}$ on the right by:

$$\begin{array}{rcl} (X \times G_{\mathbb{C}}) \times G_{\mathbb{C}} & \longrightarrow & X \times G_{\mathbb{C}} \\ ((x,g),h) & \longmapsto & (x,gh) \end{array}$$

Let $\Omega \subset X \times G_{\mathbb{C}}$ be a right $G_{\mathbb{R}}$-invariant domain and have connected fibres of p; and $u \in C^{\infty}(\Omega)$ be a right $G_{\mathbb{R}}$-invariant function. u naturally induces a smooth function $\dot{u}(x,\psi(g))$ on $\dot{\Omega} := (id_X, \psi)(\Omega)$.

Suppose that (1) u is p.s.h on Ω, (2) $\forall x \in p(\Omega), u(x,\cdot)$ is strictly p.s.h. on $\Omega_x = \Omega \cap p^{-1}(x)$, and (3) $\dot{u}(x,\cdot)$ is exhaustive on $\dot{\Omega}_x = \psi(\Omega_x)$, then the minimum principle asserts that $v(x) = \inf\limits_{g \in \Omega_x} u(x,g)$ is C^{∞} and p.s.h. on $p(\Omega)$.

Remark. C.O. Kiselman in [14] first obtained the minimum principle when $X = \mathbb{C}^n, G_{\mathbb{C}} = \mathbb{C}^m, G_{\mathbb{R}} = Im\mathbb{C}^m$, J.J. Loeb in [18] generalized Kiselman's result to the present general case.

It's easy to construct invariant p.s.h. functions w.r.t. compact Lie group via "averaging technique". However, such a technique doesn't hold again for non compact Lie group.

Observation. Let G be a real Lie group which acts on $\mathbb{C}^n$ linearly. Let D be a Bergman hyperbolic domain which is G-invariant. Then the Bergman kernel $K_D(z,\overline{w})$ satisfies $K_D(z,\overline{z}) = K_D(g \cdot z, \overline{g \cdot z})$ for $g \in G$, when G is compact; when G is semisimple, we have $K_D(z,\overline{w}) = K_D(g \cdot z, \overline{g \cdot w})$.

Brief proof is as follows. Since G linearly act on $\mathbb{C}^n$, one has a representation $G \to GL(n,\mathbb{C})$; if G is semisimple, then the image of G must be in $SL(n,\mathbb{C})$; if G is compact, the image of G is in $U(n)$. Using the transformation formula for the Bergman kernels and noting that the determinant of the Jacobian of the map $z \to g \cdot z$ is 1 for semisimple case, and is in S^1 for compact case, then we can get the result.

We consider the following question: Let X be a Stein manifold, $G_{\mathbb{C}}$ be a connected complex reductive Lie group acting on X s.t. the action is holomorphic, $G_{\mathbb{R}}$ a connected real form of $G_{\mathbb{C}}$. Let $D \subset X$ be a $G_{\mathbb{R}}$-invariant orbit connected Stein domain, is $G_{\mathbb{C}} \cdot D$ also Stein?

When $G_{\mathbb{R}}$ is compact, the answer is positive (cf. [22]). This is a special case of Theorem 1 in the section 3.

The extended future tube conjecture is a special case of the question, where $X = \mathbb{C}^{4N}, G_{\mathbb{C}} = L_+(\mathbb{C}), G_{\mathbb{R}} = L_+^{\uparrow}, D = \tau_N^+, G_{\mathbb{C}} \cdot D = \tau_N'$

Consider $X \times G_{\mathbb{C}} \xrightarrow{\rho} X, \rho(x,g) = g^{-1} \cdot x$. Suppose that there is a suitable $G_{\mathbb{R}}$-invariant s.p.s.h. function φ on D. We have a p.s.h. function $u(x,g) = \varphi(g^{-1} \cdot x)$ on $\Omega = \rho^{-1}(D)$. Define $\psi(x) = \inf_{g \in \Omega_x} u(x,g)$ for $x \in p(\Omega)$, where $p : X \times G_{\mathbb{C}} \to X$ is given by $p(x,g) = x$, and $\Omega_x := \{g \in G_{\mathbb{C}} : (X,g) \in \Omega\}$.

In order to prove $\psi(x)$ is p.s.h. on $p(\Omega) = G_{\mathbb{C}} \cdot D$, we can use the minimum principle due to Kiselman-Loeb.

Observation. Ω_x is connected if and only if D is orbit connected.

In order to use the minimum principle, we still need to check two assumptions: (i) $u(x,\cdot)$ is s.p.s.h. on Ω_x; (ii) $\dot{u}(x,\cdot)$ is exhaustion on $\dot{\Omega}_x$, where $\dot{u}(x,\psi(g))$ is defined on $\dot{\Omega} = (id,\psi)(\Omega) \subset X \times G_{\mathbb{C}}/G_{\mathbb{R}}$ and is induced by $u, \psi : G_{\mathbb{C}} \to G_{\mathbb{C}}/G_{\mathbb{R}}, \dot{\Omega}_x = \psi(\Omega_x)$. Usually speaking, assumption (i) fails on the whole Ω. However, when X has FPIG, then the assumption (i) is fulfilled on a Zariski open subset of Ω. Let $X' := \{x \in X : G_{\mathbb{C}}x \text{ is closed, } (G_{\mathbb{C}})_x \text{ is principal and finite }\}$, then, by the slice theory, $A = X \backslash A'$ is a $G_{\mathbb{C}}$-invariant analytic subset of X. Let $D' = D \cap X', \Omega' := \rho^{-1}(D')$, then the assumption (i) is satisfied on Ω'. If the assumption (ii) is also satisfied on Ω', then we can use the minimum principle on Ω' and get that $\psi(x)$ is p.s.h. on $p(\Omega') = G_{\mathbb{C}} \cdot D \backslash A$ since $\psi(x)$ is upper semicontinuous on $G_{\mathbb{C}} \cdot D$, by the extension theorem for p.s.h. functions, $\psi(x)$ can be extended to a p.s.h. function on $G_{\mathbb{C}} \cdot D$.

If we can prove that the extended p.s.h. function is weak exhaustion, then $G_{\mathbb{C}} \cdot D$ is Stein.

As a consequence of our observations, we deduce that the general question is true for pseudoconvex pair $(G_{\mathbb{C}}, G_{\mathbb{R}})$ (i.e., there exists a $G_{\mathbb{R}}$-invariant p.s.h. function on $G_{\mathbb{C}}$ which is exhaustion on $G_{\mathbb{C}}/G_{\mathbb{R}}$(cf.[17]), which include the case when $G_{\mathbb{R}}$ is compact and nilpotent(cf.[17]). However it's pity that $(L_+(\mathbb{C}), L_+^{\uparrow})$ is not a pseudoconvex pair.

In the case of the extended future tube conjecture, we proved that the assumption (ii) in the minimum principle is satisfied and the constructed function is weak exhaustion. These are the main technical difficulties. We overcome them and finished our proof via a consideration of the matrix form of the conjecture and explicit calculations based on Hua's work and matrix techniques ([9,19]).

Theorem [38,39]. *The extended future tube conjecture is true.*

A.G. Sergeev posed an interesting idea to attack the mentioned question. He assumed an invariant version of Cartan's lemma: if $A \subset D$ is a $G_{\mathbb{R}}$-invariant analytic subset, $f \in \mathcal{O}(A)^{G_{\mathbb{R}}}$, then there exists an $F \in \mathcal{O}(D)^{G_{\mathbb{R}}}$ s.t. $F|A = f$. If this is the case, we can prove that $\pi(D)$ is Stein in $X//G_{\mathbb{C}}$. In order to prove it, it's sufficient to prove $\pi(D)$ is holomorphically convex. Let $\{y_n\} \subset \pi(D)$ be an arbitrary discrete set. Then $\{\pi^{-1}(y_n)\} \cap D$ is a $G_{\mathbb{R}}$-invariant analytic subset in D. By the assumption, then there exists a $G_{\mathbb{R}}$-invariant holomorhic function F on D s.t. $F|_{\pi^{-1}}(y_n) = n$. Since $\mathcal{O}(\pi(D)) \cong \mathcal{O}(D)^{G_{\mathbb{R}}}$, then we get a holomorphic function on $\pi(D)$ which is unbounded on $\{y_n\}$. This means that $\pi(D)$ is holomorphically convex, and then $\pi^{-1}(\pi(D))$ is also Stein. When $\pi^{-1}(\pi(D)) = G_{\mathbb{C}} \cdot D$, i.e., $G_{\mathbb{C}} \cdot D$ is π-saturated, then $G_{\mathbb{C}} \cdot D$ is Stein.

It seems to be hard to prove directly the invariant version of Cartan's lemma for a noncompact Lie group $G_{\mathbb{R}}$, although it's trivially the case for a compact Lie group. Actually, we have the following:

Proposition ([41]). *Suppose, furthermore, $G_{\mathbb{C}} \cdot D$ is π-saturated. Then the invariant version of Cartan's lemma holds if and only if $G_{\mathbb{C}} \cdot D$ is Stein.*

However, we recently observed that it should be possible to directly give an answer to the above question based on L^2-methods and group actions.

References

[1] E. Casadio Tarabushi, S. Trapani: Envelopes of holomorphy of Hartogs and circular domains. *Pacific J. Math.* 149 (1991), no. 2, 231–249.

[2] E. Casadio Tarabushi, S. Trapani: Construction of envelopes of holomorphy for some classes of special domains. *J. Geom. Anal.* 4 (1994), no. 1, 1–21.

[3] G. Coeuré, J.J. Loeb: Univalence de certaines enveloppes d'holomorphie. (French) *C. R. Acad. Sci. Paris* Sér. Math. 302 (1986), no. 2, 59–61.

[4] G. Fels, L. Geati: Invariant domains in complex symmetric spaces, *J.rein und angew. Math.* 454 (1994), 97-118.

[5] H. Grauert: *Selected papers*, With commentary by Y. T. Siu et al. Springer-Verlag, 1994.

[6] H. Grauert, R. Remmert: *Coherent analytic sheaves*, Springer-Verlag, Berlin Heidelberg, 1984.

[7] H. Grauert, R. Remmert: *Theory of Stein spaces,* Grundl. 236, Springer-Verlag, 1979.

[8] L. Hörmander: *Introduction to complex analytic in several variables*, third revised ed., North-Holland Mathematical Library, Vol.7, North-Halland, Amsterdom, 1991.

[9] L.-K. Hua: *Harmonic analysis of functions of several complex variables in the classical domains,* (in Chinese) Science Press, Beijing, 1958; English translation, Amer.Math.Soc., Providence, RI, 1963.

[10] M. Jarnicki, P. Pflug: *Extension of holomorphic functions.* de Gruyter Expositions in Mathematics, 34. Walter de Gruyter Co., Berlin, 2000.

[11] M. Jarnicki, P.Pflug: On the extended tube conjecture. *Manuscripta Math.*, 89 (1996), no. 4, 461–470.
[12] R. Jost: *The general theory of quantized fields.* Amer. Math. Soc., Providence, R. I., 1965.
[13] David Kazhdan: Introduction to QFT . in *Quantum fields and strings: a course for mathematicians.* Vol. 1, 2. pp.377–418 American Mathematical Society, Providence, RI; Institute for Advanced Study (IAS), Princeton, NJ, 1999.
[14] C.O. Kiselman: The partial Legendre transformation for plurisubharmonic functions. *Invent. Math.* 49, 137–148 (1978).
[15] C.O. Kiselman: Plurisubharmonic functions and potential theory in several complex variables. in *"Developments of Mathematics 1950-2000"*, ed. by J.-P. Pier, pp.655-714, Birkhauser-verlag, 2000.
[16] M. Lassalle: Séries de Laurent des fonctions holomorphes dans la complexification d'un espace symétrique compact. (French) *Ann. Sci. école Norm. Sup.* (4) 11 (1978), no. 2, 167–210.
[17] J.J. Loeb: Pseudo-convexite des ouverts invariants et convexite geodesique dans certains espaces symetriques.*Sém. Lelong-Skoda, Lect. Notes in Math.*1198, 172–190.
[18] J.J. Loeb: Action d'une forme réele sur un groupe de Lie complexe. *Ann. Inst. Fourier,* fassc. 4, t.35(1985).
[19] Qikeng Lu: *Classical manifolds and classical domains* (in Chinese), Shanghai Scientific and Technical Press, 1963.
[20] D. Luna: Slices étales. *Bull. Soc. Math. France.* Mem 33, 81–105(1973).
[21] A.G. Sergeev: Around the extended future tube conjecture, *Lect. Notes in Math.*, v.1574, Springer-Verlag, 1994.
[22] A. G. Sergeev, P. Heinzner: The extended matrix disk is a domain of holomorphy. *Math. USSR Izvestija,* Vol.38(1992), no.3.
[23] A. G. Sergeev, V.S. Vladimirov: Complex analysis in the future tube, in *Encyclopedia of Math. Sci., Vol.8 (Several Complex Variables, II)*, Springer-Verlag, 1994.
[24] A.G. Sergeev, X.Y. Zhou: On invariant domains of holomorphy (in Russian). *Proc. of Steklov Math. Institute,* Tom 203, 159–172, 1994.
[25] A.G. Sergeev, X.Y. Zhou: Extended future tube conjecture. *Proc. of Steklov Math. Institute,* Tom 228, 32–51, 2000.
[26] Y.-T. Siu: Pseudoconvexity and the problem of Levi. *Bull. Amer. Math. Soc.* 84 (1978), no. 4, 481–512.
[27] D. M. Snow: Reductive Group Actions on Stein Spaces. *Math. Ann.* 259, 79–97 (1982).
[28] R. F. Streater, A.S. Wightman:*PCT, Spin and statistics, and all that.* Benjamin, Reading, Mass, 1964.
[29] V. S. Vladimirov: Analytic functions of several complex variables and quantum field theory. *Proc. of the Steklov Inst. of Math.* 1978, Issue 1, 69–81.
[30] V. S. Vladimirov: Several complex variables in mathematical physics. *Sém. Lelong-Skoda, Lecture Notes in Math.*, Vol.919, 1982, 358–386.

[31] V. S. Vladimirov, V. V. Zharinov: Analytic methods in mathematicial physics. *Proc. of the Steklov Inst. of Math.* 1988, Issue 2, 117–137.
[32] A. S. Wightman: Quantum field theory and analytic function of several complex variables, *J. Indian Math. Soc.* (N.S.) 24(1960/1961), 625–677.
[33] B. I. Zav'yalov, V.B.Trushin: On the extended n-point tube,*Teoret. Mat. Fiz.* 27, 1(1976), 3–15.
[34] X.Y. Zhou: On matrix Reinhardt domains. *Math.Ann.* 287, 35–46(1990).
[35] X.Y. Zhou: On orbit convexity of certain torus invariant domain of holomorphy. *Dokl. AN SSSR,* T.322, N.2, 1992, 262–267.
[36] X.Y. Zhou: On orbit connectedness, orbit convexity, and envelopes of holomorphy.*Izvestiya Ross.Akad.Nauk*, Series Math. T.58, N.2, 1994, 196–205.
[37] X.Y. Zhou: On invariant domains in certain homogeneous spaces. *Ann. L'Inst. Fourier*, T.47, N.4, 1997, 1101–1115.
[38] X.Y. Zhou: A proof of the extended future tube conjecture(in Russian).*Izvestiya Ross.Akad.Nauk*, Series Math. T.62, N.1, 1998, 211–224.
[39] X.Y. Zhou: The extended future tube is a domain of holomorphy. *Math. Research Letters* 5, 185–190(1998).
[40] X.Y. Zhou: Quotients, invariant version of Cartan's lemma, and the minimum principle. *Proc. of first ICCM.*, 335–343, Amer. Math. Soc. and International Press, 2001.
[41] X.Y. Zhou: Invariant version of Cartan's lemma and complexicification of invariant domains (in Russian). *Dokl. Ross. Akad. Nauk,* vol.366, no.5, 1999, 608–612.

Section 9. Operator Algebras and Functional Analysis

Semyon Alesker: *Algebraic Structures on Valuations, Their Properties and Applications* 757

P. Biane: *Free Probability and Combinatorics* 765

D. Bisch: *Subfactors and Planar Algebras* 775

Liming Ge: *Free Probability, Free Entropy and Applications to von Neumann Algebras* 787

V. Lafforgue: *Banach KK-theory and the Baum-Connes Conjecture* 795

R. Latała: *On Some Inequalities for Gaussian Measures* 813

ICM 2002 · Vol. II · 757–764

Algebraic Structures on Valuations, Their Properties and Applications

Semyon Alesker*

Abstract

We describe various structures of algebraic nature on the space of continuous valuations on convex sets, their properties (like versions of Poincaré duality and hard Lefschetz theorem), and their relations and applications to integral geometry.

2000 Mathematics Subject Classification: 46, 47.
Keywords and Phrases: Valuations, Convex sets, Kinematic formulas, Reductive Lie group.

0. Introduction

The theory of continuous valuations on convex sets generalizes, in a sense, both the measure theory and the theory of the Euler characteristic. Roughly speaking one should think of a continuous valuation ϕ on a real linear space V as a finite additive measure on a class of compact nice subsets of V (say piecewise smooth submanifolds with corners) which satisfy the following additional property (instead of the usual sigma-additivity): the restriction of ϕ to the subclass of convex compact domains with smooth boundary extends by continuity to the class $\mathcal{K}(V)$ of *all* convex compact subsets of V. Here the continuity is understood in the sense of the Hausdorff metric on $\mathcal{K}(V)$. Remind that the Hausdorff metric d_H on $\mathcal{K}(V)$ depends on the choice of the Euclidean metric on V and it is defined as follows: $d_H(A,B) := \inf\{\varepsilon > 0 | A \subset (B)_\varepsilon \text{ and } B \subset (A)_\varepsilon\}$, where $(U)_\varepsilon$ denotes the ε-neighborhood of a set U. This condition of continuity turns out to be very strong restriction and has a lot of consequences on purely algebraic level. These properties will be discussed in this paper. The simplest examples of such valuations are any smooth measure on V and the Euler characteristic. Also it turns out that one of the main tools used recently in investigations of valuations is the representation theory of real reductive groups and the Beilinson-Bernstein theory of D-modules.

Now let us give the formal definition of valuation.

*Department of Mathematics, Tel Aviv University, Ramat Aviv, 69978 Tel Aviv, Israel. E-mail: semyon@post.tau.ac.il

0.1.1 Definition. a) *A function $\phi : \mathcal{K}(V) \longrightarrow \mathbb{C}$ is called a valuation if for any $K_1, K_2 \in \mathcal{K}(V)$ such that their union is also convex one has*

$$\phi(K_1 \cup K_2) = \phi(K_1) + \phi(K_2) - \phi(K_1 \cap K_2).$$

b) *A valuation ϕ is called continuous if it is continuous with respect the Hausdorff metric on $\mathcal{K}(V)$.*

The linear space of all continuous valuations on V will be denoted by $CVal(V)$. It is a Fréchet space with the topology of uniform convergence on compact subsets of $\mathcal{K}(V)$. In Section 1 we discuss its dense subspace of polynomial smooth valuations $(PVal(V))^{sm}$ (it has the topology of inductive limit of Fréchet spaces). It turns out that this space has a natural structure of associative commutative unital algebra (when the unity is the Euler characteristic). In Section 2 we discuss the space $Val(V)$ of translation invariant continuous valuations. Its dense subspace $(Val(V))^{sm}$ of so called smooth valuations is a subalgebra of $(PVal(V))^{sm}$. It has a natural grading and satisfies a version of Poincaré duality. This property follows from the Irreducibility Theorem 2.1.3 which is by itself key result in the investigation of valuations (see Subsection 2.1). Moreover *even* smooth translation invariant continuous valuations form a graded subalgebra of $(Val(V))^{sm}$ and satisfy a version of the hard Lefschetz theorem (Subsection 2.2). This property turns out to be closely related to the cosine transform problem in the (Gelfand style) integral geometry solved recently in [6]. These properties of valuations turn out to be useful to obtain new explicit classification results on valuations with additional invariance properties. The classical Hadwiger theorem describes explicitly $SO(n)$- and $O(n)$-invariant translation invariant continuous valuations on the Euclidean space $\mathbb{R}^n$. The new result is the classification of unitarily invariant translation invariant continuous valuations on the Hermitian space $\mathbb{C}^n$ (Subsection 2.3). The main application of the classification results on valuations is integral geometric formulas. Using our classification we obtain new results in (Chern style) integral geometry of real submanifolds of complex spaces (Section 3).

1. General continuous valuations

In order to study general continuous valuations let us remind the definition of *polynomial* valuation introduced by Khovanskii and Pukhlikov [14], [15].

1.1.1 Definition. *A valuation ϕ is called polynomial of degree d if for every $K \in \mathcal{K}(V)$ the function $x \mapsto \phi(K + x)$ is a polynomial on V of degree at most d.*

Note that valuations polynomial of degree 0 are called *translation invariant* valuations. Polynomial valuations have many nice combinatorial-algebraic properties ([14], [15]). Also in [1] the author have classified explicitly rotation invariant polynomial continuous valuations on a Euclidean space.

Let us denote the space of polynomial continuous valuations on V by $PVal(V)$. One has

1.1.2 Proposition ([5]). *The space $PVal(V)$ of polynomial continuous valuations is dense in the space of all continuous valuations $CVal(V)$.*

The proof of this proposition is rather simple; it is a tricky use of a form of the Peter-Weyl theorem (for the orthogonal group $O(n)$), and in particular the convexity is not used in any essential way.

Let us remind the basic definition of a smooth vector for a representation of a Lie group. Let ρ be a continuous representation of a Lie group G in a Fréchet space F. A vector $\xi \in F$ is called G-smooth if the map $g \mapsto \rho(g)\xi$ is infinitely differentiable map from G to F. It is well known the the subset F^{sm} of smooth vectors is a G-invariant linear subspace dense in F. Moreover it has a natural topology of a Fréchet space (which is stronger than that induced from F), and the representation of G is F^{sm} is continuous.

We will especially be interested in polynomial valuations which are $GL(V)$-smooth. This space will be denoted by $(PVal(V))^{sm}$.

Example. Let μ be a measure on V with a polynomial density with respect to the Lebesgue measure. Let $A \in \mathcal{K}(V)$ be a strictly convex compact subset with smooth boundary. Then

$$\phi(K) := \mu(K + A)$$

is a continuous polynomial smooth valuation (here $K + A := \{k + a | k \in K, a \in A\}$).

Let us denote by $\mathcal{G}(V)$ the linear space of valuations on V which are finite linear combinations of valuations from the previous example. It can be shown (using Irreducibility Theorem 2.1.3) that $\mathcal{G}(V)$ is dense in $(PVal(V))^{sm}$. Let W be another linear real vector space. Let us define the exterior product $\phi \boxtimes \psi \in \mathcal{G}(V \times W)$ of two valuations $\phi \in \mathcal{G}(V)$, $\psi \in \mathcal{G}(W)$. Let $\phi(K) = \sum_i \mu_i(K + A_i)$, $\psi(L) = \sum_j \nu_j(L + A_j)$. Define

$$(\phi \boxtimes \psi)(M) := \sum_{i,j} (\mu_i \boxtimes \nu_j)(M + (A_i \times \{0\}) + (\{0\} \times B_j)),$$

where $\mu_i \boxtimes \nu_j$ denotes the usual product measure.

1.1.3 Proposition ([5]). *For $\phi \in \mathcal{G}(V)$, $\psi \in \mathcal{G}(W)$ their exterior product $\phi \boxtimes \psi \in \mathcal{G}(V \times W)$ is well defined; it is bilinear with respect to each argument. Moreover*

$$(\phi \boxtimes \psi) \boxtimes \eta = \phi \boxtimes (\psi \boxtimes \eta).$$

Now let us define a product on $\mathcal{G}(V)$. Let $\Delta : V \hookrightarrow V \times V$ denote the diagonal imbedding. For $\phi, \psi \in \mathcal{G}(V)$ let

$$\phi \cdot \psi := \Delta^*(\phi \boxtimes \psi),$$

where Δ^* denotes the restriction of a valuation on $V \times V$ to the diagonal.

1.1.4 Proposition ([5]). *The above defined multiplication uniquely extends by continuity to $(PVal(V))^{sm}$. Then $(PVal(V))^{sm}$ becomes an associative commutative unital algebra where the unit is the Euler characteristic χ.*

2. Translation invariant continuous valuations

For a linear finite dimensional real vector space V let us denote by $Val(V)$ the space of *translation invariant* continuous valuations on V. This is a Fréchet space with respect to the topology of uniform convergence on compact subsets of $\mathcal{K}(V)$. In this section we will discuss properties of this space.

2.1. Irreducibility theorem and Poincaré duality

It was shown by P. McMullen [17] that the space $Val(V)$ of translation invariant continuous valuations on V has a natural grading given by the degree of homogeneity of valuations. Let us formulate this more precisely.

2.1.1 Definition. *A valuation ϕ is called homogeneous of degree k if for every convex compact set K and for every scalar $\lambda > 0$*

$$\phi(\lambda K) = \lambda^k \phi(K).$$

Let us denote by $Val_k(V)$ the space of translation invariant continuous valuations homogeneous of degree k.

2.1.2 Theorem (McMullen [17]).

$$Val(V) = \bigoplus_{k=0}^{n} Val_k(V),$$

where $n = \dim V$.

Note in particular that the degree of homogeneity is an integer between 0 and $n = \dim V$. It is known that $Val_0(V)$ is one-dimensional and is spanned by the Euler characteristic χ, and $Val_n(V)$ is also one-dimensional and is spanned by a Lebesgue measure [10]. The space $Val_n(V)$ is also denoted by $Dens(V)$ (the space of densities on V). One has further decomposition with respect to parity:

$$Val_k(V) = Val_k^{ev}(V) \oplus Val_k^{odd}(V),$$

where $Val_k^{ev}(V)$ is the subspace of even valuations (ϕ is called even if $\phi(-K) = \phi(K)$ for every $K \in \mathcal{K}(V)$), and $Val_k^{odd}(V)$ is the subspace of odd valuations (ϕ is called odd if $\phi(-K) = -\phi(K)$ for every $K \in \mathcal{K}(V)$). The Irreducibility Theorem is as follows.

2.1.3 Theorem ([3],[2]). *The natural representation of the group $GL(V)$ on each space $Val_k^{ev}(V)$ and $Val_k^{odd}(V)$ is irreducible.*

This theorem is the main tool in further investigations of valuations and classification of them (see Subsection 2.3). This immediately implies so called McMullen's conjecture [18]. Its proof is heavily based on the use of the representation theory of real reductive groups and the Beilinson-Bernstein theory of D-modules. Another key tool in the proof of this result is the Klain-Schneider characterization of simple translation invariant continuous valuations [12], [20].

By the results of Section 1 $(Val(V))^{sm}$ is a subalgebra of $(PVal(V))^{sm}$. It is easy to see that the algebra structure is compatible with the grading, namely

$$(Val_i(V))^{sm} \otimes (Val_j(V))^{sm} \longrightarrow (Val_{i+j}(V))^{sm}.$$

In particular we have

$$(Val_i(V))^{sm} \otimes (Val_{n-i}(V))^{sm} \longrightarrow Dens(V).$$

A version of the Poincaré duality theorem says that this is a perfect pairing. More precisely

2.1.4 Theorem ([5]). *The induced map*

$$(Val_i(V))^{sm} \longrightarrow (Val_{n-i}(V)^*)^{sm} \otimes Dens(V)$$

is an isomorphism.

2.2. Even translation invariant continuous valuations

Let us denote by $Val^{ev}(V)$ the subspace of *even* translation invariant continuous valuations. Then clearly $(Val^{ev}(V))^{sm}$ is a subalgebra of $(Val(V))^{sm}$. It turns out that it satisfies a version of the hard Lefschetz theorem which we are going to describe.

Let us fix on V a scalar product. Let D denote the unit ball with respect to this product. Let us define an operator $\Lambda : Val(V) \longrightarrow Val(V)$. For a valuation $\phi \in Val(V)$ set

$$(\Lambda\phi)(K) := \frac{d}{d\varepsilon}|_{\varepsilon=0}\phi(K + \varepsilon D).$$

(Note that by a result of P. McMullen [17] $\phi(K + \varepsilon D)$ is a polynomial in $\varepsilon > 0$ of degree at most n.) It is easy to see that Λ preserves the parity of valuations and decreases the degree of homogeneity by 1. In particular

$$\Lambda : Val_k^{ev}(V) \longrightarrow Val_{k-1}^{ev}(V).$$

The following result is a version of the hard Lefschetz theorem.

2.2.1 Theorem ([4]). *Let $k > n/2$. Then*

$$\Lambda^{2k-n} : (Val_k^{ev}(V))^{sm} \longrightarrow (Val_{n-k}^{ev}(V))^{sm}$$

is an isomorphism. In particular for $1 \leq i \leq 2k - n$ the map

$$\Lambda^i : (Val_k^{ev}(V))^{sm} \longrightarrow (Val_{k-i}^{ev}(V))^{sm}$$

is injective.

Note that the proof of this result is based on the solution of the cosine transform problem due to J. Bernstein and the author [6], which is the problem from (Gelfand style) integral geometry motivated by stochastic geometry and going back to G. Matheron [16].

2.3. Valuations invariant under a group

Let G be a subgroup of $GL(V)$. Let us denote by $Val^G(V)$ the space of G-invariant translations invariant continuous valuations. From the results of [2] and [4] follows the following result.

2.3.1 Theorem. *Let G be a compact subgroup of $GL(V)$ acting transitively on the unit sphere. Then $Val^G(V)$ is a finite dimensional graded subalgebra of $(Val(V))^{sm}$. It satisfies the Poincaré duality, and if $-Id \in G$ it satisfies the hard Lefschetz theorem.*

It turns out that $Val^G(V)$ can be described explicitly (as a vector space) for $G = SO(n)$, $O(n)$, and $U(n)$. In the first two cases it is the classical theorem of Hadwiger [10], the last case is new (see [4]). In order to state these results we have to introduce first sufficiently many examples.

Let Ω be a compact domain in a Euclidean space V with a smooth boundary $\partial\Omega$. Let $n = \dim V$. For any point $s \in \partial\Omega$ let $k_1(s), \dots, k_{n-1}(s)$ denote the principal curvatures at s. For $0 \leq i \leq n-1$ define

$$V_i(\Omega) := \frac{1}{n}\binom{n-1}{n-1-i}^{-1} \int_{\partial\Omega} \{k_{j_1}, \dots, k_{j_{n-1-i}}\} d\sigma,$$

where $\{k_{j_1}, \dots, k_{j_{n-1-i}}\}$ denotes the $(n-1-i)$-th elementary symmetric polynomial in the principal curvatures, $d\sigma$ is the measure induced on $\partial\Omega$ by the Euclidean structure. It is well known that V_i (uniquely) extends by continuity in the Hausdorff metric to $\mathcal{K}(V)$. Define also $V_n(\Omega) := vol(\Omega)$. Note that V_0 is proportional to the Euler characteristic χ. It is well known that $V_0, V_1, \dots, V_n$ belong to $Val^{O(n)}(V)$. It is easy to see that V_k is homogeneous of degree k. The famous result of Hadwiger says

2.3.2 Theorem (Hadwiger, [10]). *Let V be n-dimensional Euclidean space. The valuations $V_0, V_1, \dots, V_n$ form a basis of $Val^{SO(n)}(V) (= Val^{O(n)}(V))$.*

Now let us describe unitarily invariant valuations on a Hermitian space. Let W be a Hermitian space, i.e. a complex vector space equipped with a Hermitian scalar product. Let $m := \dim_{\mathbb{C}} W$ (thus $\dim_{\mathbb{R}} W = 2m$). For every non-negative integers p and k such that $2p \leq k \leq 2m$ let us introduce the following valuations:

$$U_{k,p}(K) = \int_{E \in \mathbb{C}\mathcal{A}Gr_{m-p}} V_{k-2p}(K \cap E) \cdot dE.$$

Then $U_{k,p} \in Val_k^{U(m)}(W)$.

2.3.3 Theorem ([4]). *Let W be a Hermitian vector space of complex dimension m. The valuations $U_{k,p}$ with $0 \leq p \leq \frac{min\{k, 2m-k\}}{2}$ form a basis of the space $Val_k^{U(m)}(W)$.*

It turns out that the proof of this theorem is highly indirect, and it uses everything known about even translation invariant continuous valuations including

the solution of McMullen's conjecture, cosine transform, hard Lefschetz theorem for valuations, and also results of Howe and Lee [11] on the structure of certain $GL_n(\mathbb{R})$-modules. Namely in order to describe explicitly the (finite dimensional) space of unitarily invariant valuations it is necessary to study the (infinite dimensional) $GL_{\mathbb{R}}(W)$-module $Val^{ev}(W)$.

Note that as algebra $Val^{SO(n)}(V)$ is isomorphic to $\mathbb{C}[x]/(x^{n+1}$. The algebra structure of $Val^{U(m)}(W)$ is not yet computed.

3. Applications to integral geometry

In this section we state new results from (Chern style) integral geometry of Hermitian spaces. They are obtained by the author in [4] using the classification of unitarily invariant valuations described in Subsection 2.3 of this paper. They can be considered as a generalization of the classical kinematic formulas due to Chern, Crofton, Santaló, and others (see e.g. [7], [8],[9],[13], [19]).

Let us remind first the principal kinematic formula following Chern [7]. Let $ISO(n)$ denote the group of affine isometries of the Euclidean space $\mathbb{R}^n$. Let Ω_1, Ω_2 be compact domains with smooth boundary in $\mathbb{R}^n$. Assume also that $\Omega_1 \cap U(\Omega_2)$ has finitely many components for all $U \in ISO(n)$.

3.1.1 Theorem ([7]).

$$\int_{U\in ISO(n)} \chi(\Omega_1 \cap U(\Omega_2))dU = \sum_{k=0}^{n} \kappa_k V_k(\Omega_1)V_{n-k}(\Omega_2),$$

where κ_k are constants depending on k and n only which can be written down explicitly.

For the explicit form of the constants κ_k we refer to [7] or [19], Ch.15 §4.

Let us return back to the Hermitian situation. Let $IU(m)$ denote the group of affine isometries of the Hermitian space $\mathbb{C}^m$ preserving the complex structure (then $IU(m)$ is isomorphic to $\mathbb{C}^m \rtimes U(m)$). Let Ω_1, Ω_2 be compact domains with smooth boundary in $\mathbb{C}^m$ such that $\Omega_1 \cap U(\Omega_2)$ has finitely many components for all $U \in IU(m)$. The new result is

3.1.2 Theorem ([4]).

$$\int_{U\in IU(m)} \chi(\Omega_1 \cap U(\Omega_2))dU = \sum_{k_1+k_2=2m} \sum_{p_1,p_2} \kappa(k_1,k_2,p_1,p_2)U_{k_1,p_1}(\Omega_1)U_{k_2,p_2}(\Omega_2),$$

where the inner sum runs over $0 \le p_i \le k_i/2$, $i = 1,2$, and $\kappa(k_1,k_2,p_1,p_2)$ are certain constants depending on m,k_1,k_2,p_1,p_2 only.

Remark. We could compute explicitly the constants $\kappa(k_1,k_2,p_1,p_2)$ only in $\mathbb{C}^2$.

For more integral geometric formulas of this and other type for real domains in $\mathbb{C}^m$ we refer to [4].

References

[1] Alesker, Semyon; Continuous rotation invariant valuations on convex sets. Ann. of Math. (2) 149 (1999), no. 3, 977–1005.

[2] Alesker, Semyon; On P. McMullen's conjecture on translation invariant valuations. Adv. Math. 155 (2000), no. 2, 239–263.

[3] Alesker, Semyon; Description of translation invariant valuations on convex sets with solution of P. McMullen's conjecture. Geom. Funct. Anal. 11 (2001), no. 2, 244–272.

[4] Alesker, Semyon; Hard Lefschetz theorem for valuations, unitarily invariant valuations, and complex integral geometry. in preparation.

[5] Alesker, Semyon; The multiplicative structure on valuations on convex sets. in preparation.

[6] Alesker, Semyon; Bernstein, Joseph; Range characterization of the cosine transform on higher Grassmannians. math.MG/0111031

[7] Chern, Shiing-shen; On the kinematic formula in integral geometry. J. Math. Mech. 16 (1966), 101–118.

[8] Chern, Shiing-shen; On the kinematic formula in the Euclidean space of n dimensions. Amer. J. Math. 74, (1952). 227–236.

[9] Griffiths, Phillip A.; Complex differential and integral geometry and curvature integrals associated to singularities of complex analytic varieties. Duke Math. J. 45 (1978), no. 3, 427–512.

[10] Hadwiger, Hugo; Vorlesungen über Inhalt, Oberfläche und Isoperimetrie. Springer-Verlag, Berlin-Göttingen-Heidelberg 1957.

[11] Howe, Roger; Lee, Soo Teck; Degenerate principal series representations of $GL_n(\mathbb{C})$ and $GL_n(\mathbb{R})$. J. Funct. Anal. 166 (1999), no. 2, 244–309.

[12] Klain, Daniel A.; A short proof of Hadwiger's characterization theorem. Mathematika 42 (1995), no. 2, 329–339.

[13] Klain, Daniel A.; Rota, Gian-Carlo; Introduction to geometric probability. Lezioni Lincee. [Lincei Lectures] Cambridge University Press, Cambridge, 1997.

[14] Pukhlikov, A. V.; Khovanskii, A. G.; Finitely additive measures of virtual polyhedra. (Russian) Algebra i Analiz 4 (1992), no. 2, 161–185; translation in St. Petersburg Math. J. 4 (1993), no. 2, 337–356.

[15] Pukhlikov, A. V.; Khovanskii, A. G.; The Riemann-Roch theorem for integrals and sums of quasipolynomials on virtual polytopes. (Russian) Algebra i Analiz 4 (1992), no. 4, 188–216; translation in St. Petersburg Math. J. 4 (1993), no. 4, 789–812.

[16] Matheron, G.; Un théorème d'unicité pour les hyperplans poissoniens. J. Appl. Probability 11 (1974), 184–189.

[17] McMullen, Peter; Valuations and Euler-type relations on certain classes of convex polytopes. Proc. London Math. Soc. (3) 35 (1977), no. 1, 113–135.

[18] McMullen, Peter; Continuous translation-invariant valuations on the space of compact convex sets. Arch. Math. (Basel) 34 (1980), no. 4, 377–384.

[19] Santaló, Luis A.; Integral geometry and geometric probability. With a foreword by Mark Kac. Encyclopedia of Mathematics and its Applications, Vol. 1. Addison-Wesley Publishing Co., Reading, Mass.-London-Amsterdam, 1976.

[20] Schneider, Rolf; Simple valuations on convex bodies. Mathematika 43 (1996), no. 1, 32–39.

ICM 2002 · Vol. II · 765–774

Free Probability and Combinatorics

P. Biane*

Abstract

A combinatorial approach to free probability theory has been developped by Roland Speicher, based on the notion of noncrossing cumulants, a free analogue of the classical theory of cumulants in probability theory. We review this theory, and explain the connections between free probability theory and random matrices. We relate noncrossing cumulants to classical cumulants and also to characters of large symmetric groups. Finally we give applications to the asymptotics of representations of symmetric groups, specifically to the Littlewood-Richardson rule.

2000 Mathematics Subject Classification: 46L54, 05E10, 60B15.
Keywords and Phrases: Free probability, Symmetric group, Noncrossing partitions.

1. Introduction

Free probability has been introduced by D. Voiculescu [21] as a means of studying the group von Neumann algebras of free groups, using probabilistic techniques. His theory has become very successful when he discovered a deep relation with the theory of random matrices, and solved some old questions in operator algebra, see [4], [7], [24] for an overview. A purely combinatorial approach to Voiculescu's definition of freeness has been given by R. Speicher [19], [20], building on G. C. Rota's [16] approach to classical probability. It is based on the notion of noncrossing partitions, also known as "planar diagrams" in quantum field theory, and provides unifying concepts for many computations in free probability. Noncrossing partitions turn out to be connected with the geometry of the symmetric group, and this leads to some new understanding of the asymptotic behaviour of the characters and representations of large symmetric groups. Our aim is to survey these results, we shall start with the basic definition of freeness, then explain its connection to random matrix theory. In the third section we review Speicher's theory. In the fourth section we show how noncrossing cumulants arise naturally in connection

*Département de Mathématiques et Applications, École Normale Supérieure, 45 rue d'Ulm 75005 Paris, France. E-mail: Philippe.Biane@ens.fr

with classical cumulants associated with random matrices, and with characters of symmetric groups. Finally in section 5 we explain the asymptotic behaviour of representations of symmetric groups in terms of free probability concepts.

2. Freeness and random matrices

The usual framework for free probability is a von Neumann algebra A, equipped with a faithful, tracial, normal state τ. To any self-adjoint element $X \in A$ one can associate its distribution, the probability measure on the real line, uniquely determined by the identity $\tau(X^n) = \int_{\mathbf{R}} x^n \mu(dx)$ for all $n \geq 1$. This makes it natural to think of the elements of A as noncommutative random variables, and of τ as an expectation map, and one usually calls noncommutative probability space such a pair (A, τ). Although a great deal of the theory, especially the combinatorial side, can be developped in a purely algebraic way, assuming only that A is a complex algebra with unit, and τ a complex linear functional, we shall stick to the von Neumann framework in the present exposition.

Given (A, τ), one considers a family $\{A_i; i \in I\}$ of von Neumann subalgebras. This family is called a *free family* if the following holds: for any $k \geq 1$ and k-tuple $a_1, \ldots, a_k \in A$ such that

- each a_j belongs to some algebra A_{i_j}, with $i_1 \neq i_2, i_2 \neq i_3, \ldots, i_{k-1} \neq i_k$,
- $\tau(a_j) = 0$ for all j,

one has $\tau(a_1 \ldots a_k) = 0$.

Moreover, a family of elements of A is called free if the von Neumann algebras each of them generates form a free family. Freeness is a noncommutative notion analogous to the independence of σ-fields in probability theory, but which incorporates also the notion of algebraic independence.

Observe that if a_1 and a_2 are free elements in (A, τ), and one defines the centered elements $\hat{a}_i = a_i - \tau(a_i)1$ then one can conpute

$$\tau(a_1 a_2) = \tau(\hat{a}_1 \hat{a}_2) + \tau(a_1)\tau(a_2) = \tau(a_1)\tau(a_2)$$

where the freeness condition has been used to get $\tau(\hat{a}_1 \hat{a}_2) = 0$. Actually, if $\{A_i; i \in I\}$ is a free family, it is not dificult to see that one can compute the value of τ on any product of the form $a_1 \ldots a_k$, where each a_j belongs to some of the A_i's, in terms of the quantities $\tau(a_{j_1} \ldots a_{j_l})$ where all the elements $a_{j_1}, \ldots, a_{j_l}$ belong to the same subalgebra. This implies that the value of τ on the algebra generated by the family $\{A_i; i \in I\}$ is completely determined by the restrictions of τ to each of these subalgebras. However the problem of finding an explicit formula is nontrivial, and this is where combinatorics comes in. We shall describe Speicher's theory of noncrossing cumulants, which solves this problem, in the next section, but before that we explain how free probability is relevant to understand large random matrices.

Consider n random $N \times N$ matrices $X_1^{(N)}, \ldots, X_n^{(N)}$, of the form

$$X_j^{(N)} = U_j D_j^{(N)} U_j^* \tag{2.1}$$

where $D_j^{(N)}; j = 1, \ldots, n$ are diagonal, hermitian, nonrandom matrices and U_j are independent unitary random matrices, each distributed with the Haar measure on the unitary group $\mathbf{U}(N)$. In other words we have fixed the spectra of the $X_i^{(N)}$ but their eigenvectors are chosen at random. The n-tuple $X_1^{(N)}, \ldots, X_n^{(N)}$ can be recovered, up to a global unitary conjugation $X_i^{(N)} \mapsto UX_i^{(N)}U^*$, (where U does not depend on i), from its mixed moments, i.e. the set of complex numbers $\frac{1}{N}Tr(X_{i_1}^{(N)} \ldots X_{i_k}^{(N)})$ where $i_1, \ldots i_k$ are arbitrary sequences of indices in $\{1, \ldots, n\}$. In particular the spectrum of any noncommutative polynomial of the $X_i^{(N)}$ can be recovered from these data. A most remarkable fact is that if we assume that the individual moments $\frac{1}{N}Tr((X_i^{(N)})^k)$ converge as N tends to infinity, then the mixed moments $\frac{1}{N}Tr(X_{i_1}^{(N)} \ldots X_{i_k}^{(N)})$ converge in probability, and their limit is obtained by the prescriptions of free probability.

Theorem 1. *Let (A, τ) be a noncommutative probability space with free self-adjoint elements $X_1, \ldots, X_n$, satisfying $\tau(X_i^k) = \lim_{N\to\infty} \frac{1}{N}Tr((X_i^{(N)})^k)$, for all i and k, then, in probability, $\frac{1}{N}Tr(X_{i_1}^{(N)} \ldots X_{i_k}^{(N)}) \to_{N\to\infty} \tau(X_{i_1} \ldots X_{i_k})$, for all $i_1, \ldots, i_k$.*

This striking result was first proved by D. Voiculescu [23], and has lead to the resolution of many open problems about von Neumann algebras, upon which we shall not touch here.

3. Noncrossing partitions and cumulants

A partition of the set $\{1, \ldots, n\}$ is said to have a crossing if there exists a quadruple (i, j, k, l), with $1 \leq i < j < k < l \leq n$, such that i and k belong to some class of the partition and j and l belong to another class. If a partition has no crossing, it is called noncrossing. The set of all noncrossing partitions of $\{1, \ldots, n\}$ is denoted by $NC(n)$. It is a lattice for the refinement order, which seems to have been first systematically investigated in [10].

Let (A, τ) be a non-commutative probability space, then we shall define a family $R^{(n)}$ of n-multilinear forms on A, for $n \geq 1$, by the following formula

$$\tau(a_1 \ldots a_n) = \sum_{\pi \in NC(n)} R[\pi](a_1, \ldots, a_n). \tag{3.1}$$

Here, for $\pi \in NC(n)$, one has defined

$$R[\pi](a_1, \ldots, a_n) = \prod_{V \in \pi} R^{(|V|)}(a_V)$$

where $a_V = (a_{j_1}, \ldots, a_{j_k})$ if $V = \{j_1, \ldots, j_k\}$ is a class of the partition π, with $j_1 < j_2 < \ldots < j_k$ and $|V| = k$ is the number of elements of V. In particular $R[1_n] = R^{(n)}$ if 1_n is the partition with only one class. Thus one has, for $n = 3$,

$$\begin{aligned} \tau(a_1a_2a_3) &= R^{(3)}(a_1, a_2, a_3) + R^{(2)}(a_1, a_2)R^{(1)}(a_3) + R^{(2)}(a_1, a_3)R^{(1)}(a_2) \\ &\quad + R^{(2)}(a_2, a_3)R^{(1)}(a_1) + R^{(1)}(a_1)R^{(1)}(a_2)R^{(1)}(a_3). \end{aligned}$$

Observe that

$$\tau(a_1 \dots a_n) = R^{(n)}(a_1, \dots, a_n) + \text{terms involving } R^{(k)} \text{ for } k < n$$

so that the $R^{(n)}$ are well defined by (3.1) and can be computed by induction on n. They are called the noncrossing (or sometimes free) cumulant functionals on A.

The formula (3.1) can be inverted to yield

$$R^{(n)}(a_1, \dots, a_n) = \sum_{\pi \in NC(n)} Moeb([\pi, 1_n])\tau[\pi](a_1, \dots, a_n).$$

Here $\tau[\pi](a_1, \dots, a_n) = \prod_{V \in \pi} \tau(a_{j_1} \dots a_{j_k})$ where $V = \{j_1, \dots, j_k\}$ are the classes of π, and $Moeb$ is the Möbius function of the lattice $NC(n)$, see [20].

For example, one has

$$\begin{aligned} R^{(1)}(a_1) &= \tau(a_1); \qquad R^{(2)}(a_1, a_2) = \tau(a_1 a_2) - \tau(a_1)\tau(a_2); \\ R^{(3)}(a_1, a_2, a_3) &= \tau(a_1 a_2 a_3) - \tau(a_1)\tau(a_2 a_3) - \tau(a_2)\tau(a_1 a_3) \\ &\quad -\tau(a_3)\tau(a_1 a_2) + 2\tau(a_1)\tau(a_2)\tau(a_3). \end{aligned}$$

Note that when the lattice of all partitions is used instead of noncrossing partitions, then one gets the usual family of cumulants (see Rota [16]), with another Möbius function.

The connection between noncrossing cumulants and freeness is the following result from section 4 of [19].

Theorem 2. *Let $\{A_i; i \in I\}$ be a free family of subalgebras of (A, τ), and $a_1, \dots, a_n \in A$ be such that a_j belongs to some A_{i_j} for each $j \in \{1, 2, \dots, n\}$. Then one has $R^{(n)}(a_1, \dots, a_n) = 0$ if there exists some j and k with $i_j \neq i_k$.*

This result leads to an explicit expression for $\tau(a_1 \dots a_n)$, where $a_1, \dots, a_n$ is an arbitrary sequence in A, such that each a_j belongs to one of the algebras $A_i; i \in I$. By Theorem 2, in the right hand side of (3.1), the terms corresponding to partitions π having a class containing two elements j, k such that a_j and a_k belong to distinct algebras give a zero contribution. Thus we have to sum over partitions in which all $j's$ belonging to a certain block of the partition are such that a_j belongs to the same algebra. Since we can express noncrossing cumulants in terms of moments we get the formula for $\tau(a_1 \dots a_n)$ in terms of the restrictions of τ to each of the subalgebras A_i. Noncrossing cumulants are a powerful tool for making computations in free probability, see [11], [12], [13], [14], [18], for some applications. We give a simple illustration below.

Let X_1 and X_2 be two self-adjoint elements which are free, then the distribution of $X_1 + X_2$, depends only on the distributions of X_1 and X_2 and can be computed as follows. Let $R^{(n)}(X_1, \dots, X_1)$ and $R^{(n)}(X_2, \dots, X_2)$, for $n \geq 1$, be the noncrossing cumulants of X_1 and X_2, then one can expand $R^{(n)}(X_1 + X_2, \dots, X_1 + X_2)$ by multilinearity as $\sum_{i_1, \dots, i_n} R^{(n)}(X_{i_1}, \dots, X_{i_n})$ where the sum is over all sequences of 1 and 2. By Theorem 2, all terms vanish except $R^{(n)}(X_1, \dots, X_1)$ and $R^{(n)}(X_2, \dots, X_2)$. It follows that

$$R^{(n)}(X_1 + X_2, \dots, X_1 + X_2) = R^{(n)}(X_1, \dots, X_1) + R^{(n)}(X_2, \dots, X_2)$$

allowing the computation of the moments of $X_1 + X_2$, hence its distribution, in terms of the distributions of X_1 and X_2. It remains to give a compact form to the relation between moments and noncrossing cumulants. For any self-adjoint element X with distribution μ, let

$$G_X(z) = \frac{1}{z} + \sum_{k=1}^{\infty} z^{-k-1} \tau(X^k) = \int_{\mathbf{R}} \frac{1}{z-x} \mu(dx)$$

be its Cauchy transform, and let

$$K(z) = \frac{1}{z} + \sum_{k=0}^{\infty} R_k z^k$$

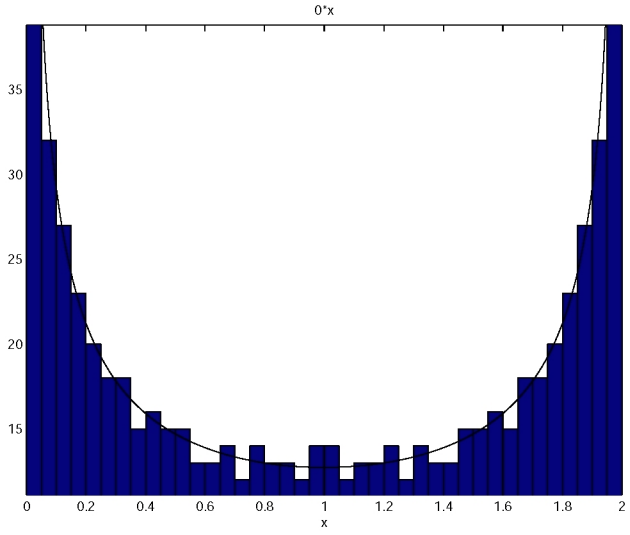
be the inverse series for composition.

Theorem 3. [19]

One has $\quad R_k = R^{(k)}(X, \ldots, X) \quad$ *for all* k.

The operation which associates to the two distributions of X_1 and X_2 the distribution of their sum is called the free convolution of measures on the real line, and was introduced by D. Voiculescu, who first considered the coefficients R_k and proved the formula for the free convolution of two measures, using very different methods [22].

Combining theorems 1 and 2, given two large random matrices of known spectra one can predict the spectral distribution of their sum, with a good accuracy and probability close to 1. It is illuminating to look at the following example. The histogram below is made of the 800 eigenvalues of a random matrix of the form $\Pi_1 + \Pi_2$ where Π_1 and Π_2 are two orthogonal projections onto some random subpaces of dimension 400 in $\mathbf{C}^{800}$, chosen independently. The curve $y = \frac{40}{\pi\sqrt{x(2-x)}}$ which corresponds to the large N limit predicted by free probability has been drawn.



4. Noncrossing cumulants, random matrices and characters of symmetric groups

Besides free probability theory, noncrossing partitions appear in several areas of mathematics. We indicate some relevant connections. The first is with the theory of map enumeration initiated by investigations of theoretical physicists in two-dimensional quantum field theory. The noncrossing partitions appear there under the guise of planar diagrams, the Feynman diagrams which dominate the matrix integrals in the large N limit. This is of course related to the fact that large matrices model free probability. We shall not discuss this further here, but refer to [26] for an accessible introduction. Another place where noncrossing partitions play a role, which is closely related to the preceding, is the geometry of the symmetric group, more precisely of its Cayley graph. Consider the (unoriented) graph whose vertex set is the symmetric group Σ_n, and such that $\{\sigma_1, \sigma_2\}$ is an edge if and only if $\sigma_1^{-1}\sigma_2$ is a transposition, i.e. this is the Cayley graph of Σ_n with respect to the generating set of all transpositions. The distance on the graph is given by

$$d(\sigma_1, \sigma_2) = n - \text{number of orbits of } \sigma_1^{-1}\sigma_2 := |\sigma_1^{-1}\sigma_2|.$$

The lattice of noncrossing partitions can be imbedded in Σ_n in the following way [10], given a noncrossing partition of $\{1, \ldots, n\}$, its image is the permutation σ such that $\sigma(i)$ is the element in the same class as i, which follows i in the cyclic order $12...n$. One can check [1] that the image of $NC(n)$ is the set of all permutations satisfying $|\sigma| + |\sigma^{-1}c| = |c|$ where c is the cyclic permutation $c(i) = i + 1\, mod(n)$, in other words, this set consists of all permutations which lie on a geodesic from the identity to c in the Cayley graph. These facts are at the heart of the connections between free probability, random matrices and symmetric groups. As an illustration we shall see how free cumulants arise from asymptotics of both random matrix theory and symmetric group representation theory.

Recall that cumulants (also called semi-invariants, see e.g. [17]) of a random variable X with moments of all orders, are the coefficients in the Taylor expansion of the logarithm of its characteristic function, i.e.

$$\log E[e^{itX}] = \sum_{n=0}^{\infty} (it)^n \frac{C_n(X)}{n!}.$$

We shall consider random variables of the following form $Y^{(N)} = NX_{1,1}^{(N)}$ where $X^{(N)} = UD^{(N)}U^*$ is a random matrix chosen as in (2.1) and $X_{1,1}^{(N)}$ is its upper left coefficient. Assume now that the moments of $X^{(N)}$ converge

$$\frac{1}{N} Tr((X^{(N)})^k) \to_{N\to\infty} \int_{\mathbf{R}} x^k \mu(dx)$$

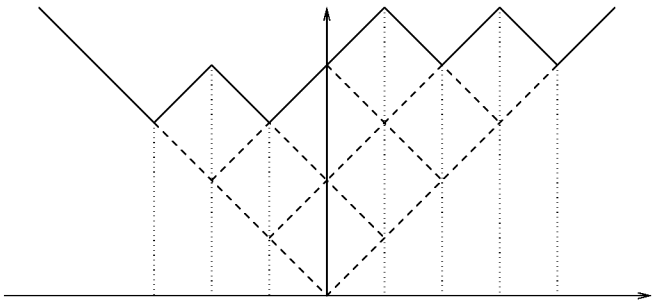
for some probability measure μ on $\mathbf{R}$, with noncrossing cumulants $R_n(\mu)$, then one has

$$\lim_{N\to\infty} \frac{1}{N^2} C_n(Y^{(N)}) = \frac{1}{n} R_n(\mu).$$

This was first observed by P. Zinn-Justin [25], a proof using representation theory has been found by B. Collins [6].

We have related noncrossing cumulants to usual cumulants via random matrix theory, we shall see that that noncrossing cumulants are also useful in evaluating characters of symmetric groups. The precise relation however is not obvious at first sight.

Let us recall a few facts about irreducible representations of symmetric groups. It is well known that they can be parametrized by Young diagrams. In the following it will be convenient to represent a Young diagram by a function $\omega : \mathbf{R} \to \mathbf{R}$ such that $\omega(x) = |x|$ for $|x|$ large enough, and ω is a piecewise affine function, with slopes ± 1, see the following picture which shows the Young diagram corresponding to the partition $8 = 3 + 2 + 2 + 1$.



Alternatively we can encode the Young diagram using the local minima and local maxima of the function ω, denoted by $x_1, \ldots, x_k$ and $y_1, \ldots, y_{k-1}$ respectively, which form two interlacing sequences of integers. These are (-3,-1,2,4) and (-2,1,3) respectively in the above picture. Associated with the Young diagram there is a unique probability measure m_ω on the real line, such that

$$\int_{\mathbf{R}} \frac{1}{z-x} m_\omega(dx) = \frac{\prod_{i=1}^{k-1}(z-y_i)}{\prod_{i=1}^{k}(z-x_i)} \quad \text{for all } z \in \mathbf{C} \setminus \mathbf{R}.$$

This probability measure is supported by the set $\{x_1, \ldots, x_k\}$ and is called the transition measure of the diagram, see [8]. Let σ denote the conjugacy class in Σ_n of a permutation with k_2 cycles of length 2, k_3 of length 3, etc.. Here $k_2, k_3, \ldots$ are fixed while we let $n \to \infty$. Denote by χ_ω the normalized character of Σ_n associated with the Young diagram ω, then the following asymptotic evaluation holds uniformly on the set of A-balanced Young diagrams, i.e. those whose longest row and longest column are less that $A\sqrt{n}$ (where A is some constant > 0),

$$\chi_\omega(\sigma) = \prod_{j=2}^{\infty} n^{-jk_j} R_{j+1}^{k_j}(\omega) + O(n^{-1-|\sigma|/2}). \tag{4.1}$$

Note that R_k is scaled by λ^k if we scale the diagram ω by a factor λ, therefore the first term in the right hand side is of order $O(n^{\sum_j (j+1)k_j/2 - \sum_j jk_j}) = O(n^{-|\sigma|/2})$,

this gives the order of magnitude of the character of a fixed conjugacy group for an A-balanced diagram.

In [2] a proof of (4.1) has been given, using in an essential way the Jucys-Murphy operators. Another proof, leading to an exact formula for characters of cycles due to S. Kerov [9], was shown to me later by A. Okounkov [15], see [5].

5. Representations of large symmetric groups

The asymptotic formula (4.1) shows in particular that irreducible characters of symmetric groups become asymptotically multiplicative i.e. for permutations with disjoint supports σ_1 and σ_2, one has

$$\chi_\omega(\sigma_1\sigma_2) = \chi_\omega(\sigma_1)\chi_\omega(\sigma_2) + O(n^{-1-|\sigma_1\sigma_2|/2}) \tag{5.1}$$

uniformly on A-balanced diagrams. Conversely, given a central, normalized, positive definite function on Σ_n, a factorization property such as (5.1) implies that the positive function is essentially an irreducible character [3]. More precisely, recall that a central normalized positive definite function ψ on Σ_n is a convex combination of normalized characters, and as such it defines a probability measure on the set of Young diagrams. For any $\varepsilon, \delta > 0$, for all n large enough, if an approximate factorization such as (5.1) holds for ψ, then there exists a curve ω, such that the measure on Young diagrams associated with ψ puts a mass larger than $1 - \delta$ on Young diagrams which lie in a neighbourhood of this curve, of width $\varepsilon\sqrt{n}$. Therefore one can say that condition (5.1) on a positive definite function implies that the representation associated with this function is approximately isotypical, i.e. almost all Young diagrams occuring in the decomposition have a shape close to a certain definite curve.

Using this fact it is possible to understand the asymptotic behaviour of several operations in representation theory. Consider for example the operation of induction. One starts with two irreducible representations of symmetric groups $\Sigma_{n_1}, \Sigma_{n_2}$, corresponding to two Young diagrams ω_1 and ω_2. One can then induce the product representation $\omega_1 \otimes \omega_2$ of $\Sigma_{n_1} \times \Sigma_{n_2}$ to $\Sigma_{n_1+n_2}$. This new representation is reducible and the multiplicities of irreducible representations can be computed using a combinatorial device, the Littlewood-Richardson rule. This rule however gives little light on the asymptotic behaviour of the multiplicities. Using the factorization-concentration result, one can prove that when n_1 and n_2 are very large, but of the same order of magnitude, then there exists a curve, which depends on ω_1 and ω_2, and such that the typical Young diagram occuring in the decomposition of the induced representation, is close to this curve. As we saw in section 4, one can associate a probability measure on the real line to any Young diagram. The description of the typical shape of Young diagram which occurs in the decomposition of the induced representation is easier if we use this correspondance between probability measures and Young diagrams, indeed the probability measure associated with the shape of the typical Young diagram corresponds to the free convolution of the two probability measures [2].

There are analogous results for the restriction of representations from large symmetric groups to smaller ones. There the corresponding operation on probability measure is called the free compression, it corresponds at the level of the large matrix approximation, to taking a random matrix with prescribed eigenvalue distribution, as in section 2, and extracting a square submatrix. Finally there are also results for Kronecker tensor products of representations. Here a central role is played by the well known Kerov-Vershik limit shape, whose associated probability measure is the semi-circle distribution with density $\frac{1}{2\pi}\sqrt{4-x^2}$ on the interval $[-2,2]$, see [2].

References

[1] P. Biane, Some properties of crossings and partitions. *Discrete Math.* 175 (1997), no. 1-3, 41–53.

[2] P. Biane, Representations of symmetric groups and free probability. *Adv. Math.* 138 (1998), no. 1, 126–181.

[3] P. Biane, Approximate factorization and concentration for characters of symmetric groups. *Internat. Math. Res. Notices* no. 4 (2001), 179–192.

[4] P. Biane, Entropie libre et algèbres d'opérateurs. *Séminaire Bourbaki* Exposé 889, Juin 2001.

[5] P. Biane, Free cumulants and characters of symmetric groups. Preprint, 2001.

[6] B. Collins, Moments and cumulants of polynomial random variables on unitary groups Preprint, 2002.

[7] L. M. Ge, these Proceedings.

[8] S. V. Kerov, Transition probabilities of continual Young diagrams and the Markov moment problem *Funct. Anal. Appl.* **27** (1993), 104–117.

[9] S.Kerov, talk at IHP, January 2000.

[10] G. Kreweras, Sur les partitions non croises d'un cycle. *Discrete Math.* 1 (1972), no. 4, 333–350.

[11] A. Nica, R. Speicher, Commutators of free random variables, *Duke Math. J.* 92 (1998), no. 3, 553–592.

[12] A. Nica, R. Speicher, On the multiplication of free N-tuples of noncommutative random variables, *Amer. J. Math.* 118 (1996), no. 4, 799–837.

[13] A. Nica, R Speicher, R-diagonal pairs—a common approach to Haar unitaries and circular elements. Free probability theory (Waterloo, ON, 1995), 149–188, Fields Inst. Commun., 12, Amer. Math. Soc., Providence, RI, 1997.

[14] A. Nica, D. Shlyakhtenko, R. Speicher, R-diagonal elements and freeness with amalgamation. *Canad. J. Math.* 53 (2001), no. 2, 355–381.

[15] A. Okounkov, private communication.

[16] G.-C. Rota, On the foundations of combinatorial theory. I. Theory of Mbius functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 2 (1964) 340–368.

[17] A. N. Shiryaev, *Probability* Graduate texts in Mathematics 95, Springer, 1991.

[18] P. Śniady, R. Speicher, Continuous family of invariant subspaces for R-diagonal operators. *Invent. Math.*, 146 (2001), no. 2, 329–363.

[19] R. Speicher, Multiplicative functions on the lattice of noncrossing partitions and free convolution. *Math. Ann.*, 298 (1994), no. 4, 611–628.

[20] R. Speicher, *Combinatorial theory of the free product with amalgamation and operator-valued free probability theory*, Mem. Amer. Math. Soc. 132 (1998), no. 627.

[21] D. Voiculescu, Symmetries of some reduced free product C^*-algebras. *Operator algebras and their connections with topology and ergodic theory* (Buşteni, 1983), Lecture Notes in Math., 1132: 556–588, Springer, Berlin-New York, 1985.

[22] D. Voiculescu, Addition of non-commuting random variables, J. Operator Theory 18 (1987) 223–235

[23] D. Voiculescu, Limit laws for random matrices and free products. *Inv. Math.* 104,(1983) 201–220.

[24] D. Voiculescu, Free probability theory: random matrices and von Neumann algebras. *Proceedings of the International Congress of Mathematicians.* Vol. 1, 2 (Zrich, 1994), 227–241, Birkhuser, Basel, 1995.

[25] P. Zinn-Justin, Universality of correlation functions of Hermitian random matrices in an external field. Comm. Math. Phys. 194 (1998), no. 3, 631–650.

[26] A. Zvonkin, Matrix integrals and map enumeration: an accessible introduction. *Math. Comput. Modelling* 26 (1997), no. 8-10, 281–304.

ICM 2002 · Vol. II · 775–785

Subfactors and Planar Algebras

D. Bisch*

Abstract

An inclusion of II_1 factors $N \subset M$ with finite Jones index gives rise to a powerful set of invariants that can be approached successfully in a number of different ways. We describe Jones' pictorial description of the standard invariant of a subfactor as a so-called planar algebra and show how this point of view leads to new structure results for subfactors.

2000 Mathematics Subject Classification: 46L37, 46L60, 82B20, 81T05.
Keywords and Phrases: Von Neumann algebras, Subfactors, Planar algebras.

1. Introduction

Abelian von Neumann algebras are simply algebras of bounded, measurable functions on a measure space. A general (non-abelian) von Neumann algebra can be viewed as an algebra of "functions" (operators) on a *non-commutative measure space.* The building blocks of what one might call *non-commutative probability spaces* are the so-called II_1 *factors* M, that is those von Neumann algebras with trivial center that are infinite dimensional and possess a distinguished tracial state (the analogue of a *non-commutative integral*). The "smallest" II_1 factor is the *hyperfinite* II_1 factor which is obtained as the closure in the weak operator topology of the canonical anti-commutation relations (CAR) algebra of quantum field theory. A II_1 factor comes always with a natural left representation on $L^2(M)$, the non-commutative L^2-space associated to M. See for instance [13].

Vaughan Jones initiated in the early 80's the *theory of subfactors* as a "Galois theory" for inclusions of II_1 factors. A *subfactor* is an inclusion of II_1 factors $N \subset M$ such that the dimension of M as left N-Hilbert module is finite. This dimension is called the *Jones index* $[M : N]$ ([19]) and one would expect by classical results of Murray and von Neumann that it takes on any real number ≥ 1. One of the early results in the theory of subfactors was Jones' spectacular *rigidity theorem* which says that this index is in fact *quantized* [19]: if $[M : N] \leq 4$, then it has to be of the form $4\cos^2 \frac{\pi}{n}$, for some $n \geq 3$. Since Jones' early work the theory of

*Department of Mathematics, Vanderbilt University, Nashville, TN 37240 and UCSB, Department of Mathematics, Santa Barbara, CA 93106, USA. E-mail: bisch@math.vanderbilt.edu

subfactors has developed into one of the most exciting and rapidly evolving areas of operator algebras with numerous applications to different areas of mathematics (e.g. knot theory with the discovery of the *Jones polynomial* [20]), quantum physics and statistical mechanics. Subfactors with finite Jones index have an amazingly rich mathematical structure and an interplay of analytical, algebraic-combinatorial and topological techniques is intrinsic to the theory.

2. Subfactors

A subfactor can be viewed as a group-like object that encodes what one might call *generalized symmetries* of the data that went into its construction. To decode this information one needs to compute the *higher relative commutants*, a system of inclusions of certain finite dimensional C*-algebras naturally associated to the subfactor. This system is an invariant of the subfactor, the so-called *standard invariant*, which contains in many natural situations precisely the same information as the subfactor itself ([30], [32], [33]). Here is one way to construct the standard invariant: If $N \subset M$ denotes an inclusion of II_1 factors with finite Jones index, and e_1 is the orthogonal projection $L^2(M) \to L^2(N)$, then we define M_1 to be the von Neumann algebra generated by M and e_1 on $L^2(M)$. M_1 is again a II_1 factor and $M \subset M_1$ has finite Jones index as well so that the previous construction can be repeated and *iterated* [19]. One obtains a tower of II_1 factors $N \subset M \subset M_1 \subset M_2 \subset \ldots$ associated to $N \subset M$, together with a remarkable sequence of projections $(e_i)_{i \geq 1}$, the so-called *Jones projections*, which satisfy the Temperley-Lieb relations and give rise to Jones' braid group representation [19], [20]. The (trace preserving) isomorphism class of the system of inclusions of (automatically finite dimensional) centralizer algebras or *higher relative commutants*

$$\begin{array}{ccccccccc} \mathbb{C} = N' \cap N & \subset & N' \cap M & \subset & N' \cap M_1 & \subset & N' \cap M_2 & \subset & \cdots \\ & & \cup & & \cup & & \cup & & \\ & & \mathbb{C} = M' \cap M & \subset & M' \cap M_1 & \subset & M' \cap M_2 & \subset & \cdots \end{array}$$

is then the standard invariant $\mathcal{G}_{N,M}$ of the subfactor $N \subset M$. Each row of inclusions is given by a sequence of Bratteli diagrams, which can in fact be reconstructed from a single, possibly infinite, bipartite graph. Hence one obtains two graphs (one for each row), the so-called *principal graphs* of $N \subset M$, which capture the inclusion structure of the above double-tower of higher relative commutants. It turns out that if M is hyperfinite and $N \subset M$ has *finite depth* (i.e. the principal graphs are finite graphs) [30], [32] or more generally if $N \subset M$ is *amenable* [33], then the standard invariant determines the subfactor. In this case the subfactor can be reconstructed from the finite dimensional data given by $\mathcal{G}_{N,M}$. In particular, subfactors of the hyperfinite II_1 factor R with index ≤ 4 are completely classified by their standard invariant and an explicit list can be given (see for instance [14], [16] or [33]). If the Jones index becomes ≥ 6 such an explicit list is out of reach as the work in [6], [11] and [12] shows: there are uncountably many non-isomorphic, irreducible infinite

depth subfactors of R with Jones index 6 and the same standard invariant! Partial lists of irreducbile subfactors with index between 4 and 6 have been obtained by different methods (see for instance [1], [5], [6], [17], [35], [36], [37], [38]), but much work remains to be done.

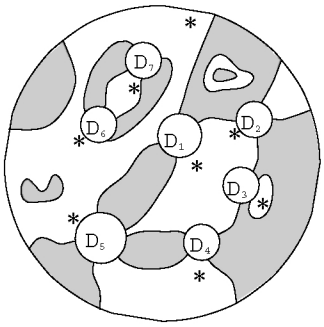
There are several distinct ways to analyze the standard invariant of a subfactor (see [2], [4], [14], [22], [30], [33]). For instance, in the bimodule approach ([13], [30], see also [4], [14], [18]) $\mathcal{G}_{N,M}$ is described as a *graded tensor category* of natural bimodules associated to the subfactor. $\mathcal{G}_{N,M}$ can thus be viewed as an *abstract system of (quantum) symmetries* of the mathematical or physical situation from which the subfactor was constructed. It is in fact a mathematical object which generalizes for instance discrete groups and representation categories of quantum groups ([37], [38]). A variety of powerful and novel techniques have been developed over the last years that make it possible to compute and understand the standard invariant of a subfactor. A key result is Popa's *abstract characterization* of the standard invariant [34]. Popa gives a set of axioms that an abstract system of inclusions of finite dimensional C*-algebras needs to satisfy in order to arise as the standard invariant of some (not necessarily hyperfinite) subfactor. This result makes it possible to analyze the structure of subfactors, which are infinite dimensional, highly non-commutative objects, by investigating the *finite dimensional* structures encoded in their standard invariants.

3. Planar algebras

Jones found in [22] a powerful formalism to handle complex computations with $\mathcal{G}_{N,M}$. He showed that the standard invariant of a subfactor has an intrinsic *planar* structure (this will be made precise below) and that certain *topological* arguments can be used to manipulate the operators living in the higher relative commutants of the subfactor. The standard invariant is a so-called *planar algebra*. To explain this notion let us first define the *planar "operad"* following [22]. Elements of the planar operad are certain classes of *planar k-tangles* which determine multilinear operations on the vector spaces underlying the higher relative commutants associated to a finite index subfactor.

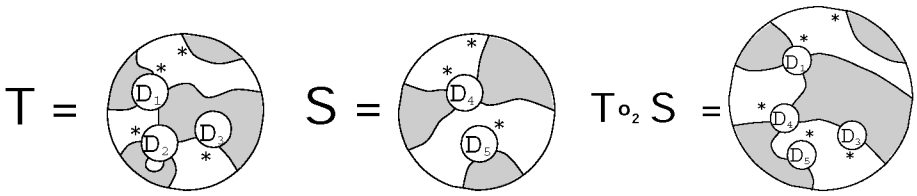
A planar k-tangle consists of the unit disk D in the complex plane together with several interior disks $D_1, D_2, \ldots, D_n$. The boundary of D is marked with $2k$ points and each D_j has $2k_j$ marked points on its boundary. These marked points are connected by strings in D, which meet the boundary of each disk transversally. We also allow (finitely many) strings which are closed curves in the interior of D. The main point is that all strings are required to be disjoint (hence *planarity*) and to lie in the complement of the interiors of the D_j's. Additional data of a planar k-tangle is a checkerboard shading of the connected components of $\overset{\circ}{D} \setminus \bigcup_{j=1}^{n} D_j$, and a choice of a white region at every D_j (which corresponds to a choice of the *first* marked point on the boundary of each D_j). The *planar operad* $\mathcal{P}$ is defined to consist of all orientation-preserving diffeomorphism classes of planar k-tangles

(for all $k \geq 0$), where the diffeomorphisms leave the boundary of D fixed but are allowed to move the interior disks. $\mathcal{P}$ becomes a *colored operad* [22] (see [28]). An example of a 4-tangle is depicted in the next figure:



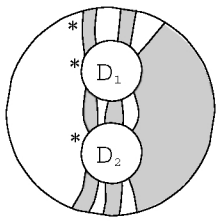
Note that there are two classes of planar 0-tangles according to the shading of the tangle near the boundary of D.

Two planar tangles $\mathcal{T}$ and $\mathcal{S}$ can be composed in a natural way if the number of boundary points of $\mathcal{S}$ matches the number of boundary points of one of the interior disks D_j of $\mathcal{T}$: To obtain the composed tangle $\mathcal{T} \circ_j \mathcal{S}$ shrink $\mathcal{S}$ and paste it inside D_j so that the shadings and marked white regions match up. Join the strings at the boundary of D_j, smooth them and erase the boundary of D_j. It is clear that this operation is well-defined (the checkerboard shading and choice of a white region at each disk avoid rotational ambiguity) and that it depends only on the isotopy class of each tangle. Note that there may be several different ways of composing two given tangles, each composition yielding potentially distinct planar tangles. An example of such a composition is given in the next figure (insert $\mathcal{S}$ in the disk D_2 of $\mathcal{T}$):



An *abstract planar algebra* is then defined to be an algebra over this planar operad ([28]). More concretely, an abstract planar algebra $\mathcal{P}$ is the disjoint union of vector spaces $\mathcal{P} = P_0^{white} \coprod P_0^{black} \coprod_{n>0} P_n$ plus a morphism from the planar operad to the (colored) operad of multilinear maps between these vector spaces. In other words a planar algebra structure on $\mathcal{P}$ is a procedure that assigns to each planar k-tangle $\mathcal{T}$ (with interior disks D_j having $2k_j$ boundary points, $1 \leq j \leq n$) a multilinear map $Z(\mathcal{T}) : P_{k_1} \times \cdots \times P_{k_n} \to P_k$ in such a way that composition of tangles is compatible with the usual composition of maps (*naturality* of composition). Note that the P_k's are automatically associative algebras since the tangle

in the next figure (drawn in the case $k = 5$) defines an associative multiplication $P_k \times P_k \to P_k$ (associativity follows from naturality of the composition).



Observe that this is a purely algebraic structure - the definition can be made for (possibly infinite dimensional) vector spaces over an arbitrary field. The key point is of course that this structure appears naturally in the theory of subfactors. In order to connect with subfactors several additional conditions will be required in the definition of a planar algebra. A *planar algebra* (or *subfactor planar algebra* to emphasize the operator algebra context) will be an abstract planar algebra such that $\dim P_k < \infty$ for all k, $\dim P_0^{white} = \dim P_0^{black} = 1$ and such that the *partition function* Z associated to the planar algebra is positive and non-degenerate. The partition function is roughly obtained as follows: If $\mathcal{T}$ is a 0-tangle, then $Z(\mathcal{T})$ is a *scalar* since it is an element in the 1-dimensional space P_0^{white} resp. P_0^{black}. Note that every planar algebra comes with two parameters $\delta_1 = Z($  $)$ and $\delta_2 = Z($  $)$, which we require to be $\neq 0$ (the inner circles are strings, not boundaries of disks!). In the case of a subfactor planar algebra we have $\delta \stackrel{def}{=} \delta_1 = \delta_2$ (which is equivalent to extremality of the subfactor [31]). In fact $\delta = [M : N]^{1/2}$ in this case. There is an intrinsic way to define an involution on the planar algebra arising from a subfactor which makes the partition function into a sesquilinear form on the standard invariant. *Positivity of the partition function* Z means then positivity of this form. Note that Z gives in particular the natural trace on the standard invariant of the subfactor. The main result of [22] is then the following theorem.

Theorem 3.1. *The standard invariant $\mathcal{G}_{N,M}$ of an extremal subfactor $N \subset M$ is a subfactor planar algebra $\mathcal{P} = (P_n)_{n \geq 0}$ with $P_n = N' \cap M_{n-1}$.*

This theorem says in particular that planar tangles *always* induce multilinear maps ("planar operations") on the standard invariant of a subfactor. As a consequence one obtains a diagrammatic formalism that can be employed to manipulate the operators in $N' \cap M_{n-1}$ and intricate calculations with these operators can be carried out using simple topological arguments. This point of view has been turned in [9], [10] into a powerful tool to prove general structure theorems for subfactors, and to analyze the rather complex combinatorial structure of the standard invariant of a subfactor. It has led to a generators and relations approach to subfactors. See also [23], [24] for more on this.

The two most fundamental examples of subfactor planar algebras are the *Temperley-Lieb systems* of [19] (see also [22]) and the *Fuss-Catalan systems* of [7] (see section 4).

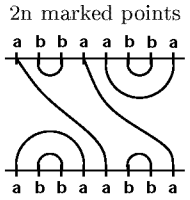
Observe that by construction planar algebras are closely related to invariants for graphs, knots and links and to the pictorial formalism commonly used in the theory of integrable lattice models in statistical mechanics.

4. Fuss-Catalan algebras

Jones and I discovered in [7] a new hierarchy of finite dimensional algebras, which arise as the higher relative commutants of subfactors when intermediate subfactors are present. These algebras have a number of interesting combinatorial properties and they have recently been used to construct new integrable lattice models and new solutions of the Yang-Baxter equation ([15], [29]).

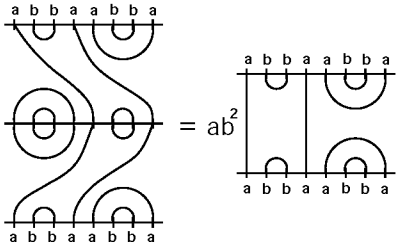
We show in [7] that a chain of $k-1$ intermediate subfactors $N \subset P_1 \subset P_2 \subset \ldots P_{k-1} \subset M$ leads to a tower of algebras $\left(FC_n(a_1,\ldots,a_k)\right)_{n\geq 0}$, which depend on k complex parameters $a_1,\ldots,a_k$. The dimensions of these algebras are given by the generalized Catalan numbers or *Fuss-Catalan numbers* $\frac{1}{kn+1}\binom{(k+1)n}{n}$ and we therefore call these algebras the *Fuss-Catalan algebras.* If no intermediate subfactor is present, i.e. $P_i = N$ or $P_i = M$ for all i, then one finds the well-known *Temperley-Lieb algebras* (case $k=1$) [19]. The additional symmetry coming from the intermediate subfactor is captured *completely* by these new algebras and it is proved in [7] (see also [8]) that they constitute the minimal symmetry present whenever an intermediate subfactor occurs. See also [26].

Let us explain in more detail what happens in the case of just one intermediate subfactor. We consider $N \subset P \subset M$, an inclusion of II_1 factors with finite Jones index, and construct the associated tower of of II_1 factors as in section 2. One obtains an inclusion of II_1 factors $N \subset P \subset M \overset{p_1}{\subset} P_1 \overset{e_1}{\subset} M_1 \overset{p_2}{\subset} P_2 \overset{e_2}{\subset} M_2 \subset \ldots$, where the p_i's are the orthogonal projections from $L^2(M_{i-1})$ onto $L^2(P_{i-1})$ ($P_0 = P$, $M_0 = M$) and the intermediate subfactors P_i are the von Neumann algebras generated by M_{i-1} and p_i. The algebra $\mathrm{IA}_n(\alpha,\beta) \overset{\mathrm{def}}{=} \mathrm{Alg}(1, e_1,\ldots,e_{n-1},p_1,\ldots,p_{n-1})$, generated by the e_i's and the p_i's, is a subalgebra of $N' \cap M_{n-1}$. It can be shown to depend only on the two indices $\alpha = [P:N]$ and $\beta = [M:P]$, and *not* on the particular position of P in $N \subset M$. The projections e_i and p_j satisfy again some rather nice commutation relations (see [7] for details). In order to describe the structure of these algebras let us for the moment consider the complex vector space $FC_n(a,b)$, spanned by labelled, planar diagrams of the form

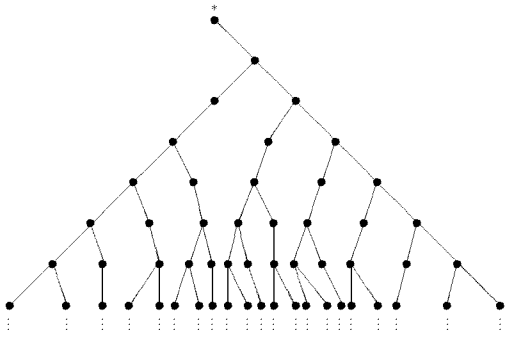


where a, $b \in \mathbb{C}\backslash\{0\}$ are fixed. There is a natural multiplication of these diagrams, which makes $FC_n(a,b)$ into an associative algebra (see [25]). To obtain $D_1 \cdot D_2$ put

the basis diagram D_1 on top of D_2 so that the labelling matches, remove the middle bar and all closed loops. Multiply the resulting diagram with factors of a resp. b according to the number of removed a-loops resp. b-loops. An example is depicted in the next figure.



Counting diagrams shows that $\dim FC_n(a,b) = \frac{1}{2n+1}\binom{3n}{n}$, the n-th Fuss-Catalan number [7]. Clearly $FC_n(a,b)$ embeds as a subalgebra of $FC_{n+1}(a,b)$ by adding two vertical through strings to the right of each basis diagram of $FC_n(a,b)$. A diagrammatic technique, called the *middle pattern analysis* in [7], can be used to compute the structure of these algebras completely in the semi-simple case. One obtains that the structure of the tower $FC_1(a,b) \subset FC_2(a,b) \subset \ldots$ of Fuss-Catalan algebras is given by the *Fibonacci graph* [7].



The algebras $\mathrm{IA}_n(\alpha,\beta)$ that we are interested in can then be shown to be isomorphic to $FC_n(a,b)$, where $\alpha = a^2$, $\beta = b^2$, if the indices α and β are generic, i.e. > 4. In the non-generic case $\mathrm{IA}_n(\alpha,\beta)$ is a certain quotient of $FC_n(a,b)$ (see [7] for the details).

There is a natural *2-parameter Markov trace* on the Fuss-Catalan algebras and the trace weights are calculated explicitly in [7]. In the special case of the Temperley-Lieb algebras this Markov trace is the one discovered by Jones in [19]. The Fuss-Catalan tower together with this Markov trace satisfies Popa's axioms in [34] and hence, one can conclude from [34] that for every pair (α,β) of possible Jones indices, there is a subfactor whose standard invariant is given precisely by the corresponding Fuss-Catalan system $\left(FC_n(\sqrt{\alpha},\sqrt{\beta})\right)_{n\geq 0}$. One obtains in this way uncountably many new subfactors. A complete set of generators and relations for the Fuss-Catalan algebras is also determined in [7].

It should be evident that the Fuss-Catalan algebras can be viewed as planar algebras generated by a single element in $P_2 = N' \cap M_1$, namely by the Jones projection p_1 onto the intermediate subfactor. This projection can be characterized abstractly [3] and it satisfies a remarkable exchange relation ([9], [27]), which plays an important role in the work described in the next section.

5. Singly generated planar algebras

Any subset S of a planar algebra $\mathcal{P}$ generates a planar subalgebra as the smallest graded vector space containing S and closed under planar operations. From this point of view the simplest subfactors will be those whose planar algebra is generated by the fewest elements satisfying the simplest relations, while the index may be arbitrarily large. If S is empty we obtain the Temperley-Lieb algebra. The next most complicated planar algebras after Temperley-Lieb should be those generated by a single element R which is in the k-graded subspace P_k for some $k > 0$. We call such an element a *k-box.* In [22] the planar algebra generated by a single 1-box was completely analyzed so the next case is that of a planar algebra generated by a single 2-box. This means that the dimension of P_2 is at least 3 so the first case to try to understand is when $\dim P_3 = 3$. This dimension condition by itself imposes many relations on $\mathcal{P}$ but probably not enough to make a complete enumeration a realistic goal. However, if one imposes $\dim P_3 \leq 15$, then apart from a degenerate case, this forces enough relations to reduce the number of variables governing the planar algebra structure to be finite in number ([22], see also [9]). It seems therefore reasonable to try to find all subfactor planar algebras $\mathcal{P}$ generated by a single element in P_2 subject to the two restrictions $\dim P_2 = 3$ and $\dim P_3 = d$ with $d \leq 15$.

In [9] we solved this problem when $d \leq 12$. In fact, using planar algebra techniques we prove a much more general structure theorem for subfactors.

Theorem 5.1. *Let $N \subset M$ be an inclusion of II_1 factors with $3 < [M : N] < \infty$. Suppose that $\dim N' \cap M_1 = 3$ and that $N' \cap M_2$ is abelian modulo the basic construction ideal $(N' \cap M_1)e_2(N' \cap M_1)$. Then there is an intermediate subfactor P of $N \subset M$, $P \neq N$, M. In particular $Jx^*J = x$ for all $x \in N' \cap M_1$.*

The proof uses in a crucial way the abstract characterization of the intermediate subfactor projection in [3] and planar algebra techniques developed in [22] and [9]. It implies the following classification result.

Theorem 5.2. *If $\mathcal{P}$ is a subfactor planar algebra generated by a 3-dimensional P_2, subject to the condition $\dim P_3 \leq 12$, then it must be one of the following:*

a) *If $\dim P_3 = 9$, then it is the planar algebra associated to the index 3 subfactor $M^{\mathbb{Z}_3} \subset M$.*
b) *If $\dim P_3 = 10$, then it is the D_∞ planar algebra (a special FC planar algebra).*
c) *If $\dim P_3 = 11$ or 12, then it is one of the FC planar algebras.*

The dimension conditions imply that a subfactor whose standard invariant is a planar algebra of the form b) or c) satisfies the hypothesis of Theorem 5.1 and hence must have an intermediate subfactor. Since the Fuss-Catalan planar algebra is the minimal symmetry associated to an intermediate subfactor it then follows easily that the planar algebra has to be one of these.

It is quite natural to expect that increasing the dimension of P_3 should result in a larger number of examples of planar algebras since there are more a priori undetermined structure constants in the action of planar tangles on $\mathcal{P}$. Thus the result in [10] that there is a *single* subfactor planar algebra satisfying the above restrictions with $d = 13$ is a complete surprise. The planar algebra which arises is that of a subfactor obtained as follows. Take an outer action of the dihedral group D_5 on a type II_1 factor R and let M be the crossed product $R \rtimes D_5$ and N be the subfactor $R \rtimes \mathbb{Z}_2$. This particular subfactor has played a significant role in the development of subfactors and relations with knot theory and statistical mechanics. In [21] it was noted that there is a solvable statisitical mechanical model associated with it and that it corresponds to an evaluation of the Kauffman polynomial invariant of a link. We prove in [10] the following

Theorem 5.3. *Let $\mathcal{P} = (P_k)_{k\geq 0}$ be a subfactor planar algebra generated by a non-trivial element in P_2 (i.e. an element not contained in the Temperley-Lieb subalgebra of P_2) subject to the conditions $\dim P_2 = 3$ and $\dim P_3 = 13$. Then $\mathcal{P}$ is the standard invariant of the crossed product subfactor $R \rtimes \mathbb{Z}_2 \subset R \rtimes D_5$. Thus there is precisely one subfactor planar algebra $\mathcal{P}$ subject to the above conditions.*

Note that this subfactor can be viewed as a Birman-Murakami-Wenzl subfactor (associated to the quantum group of $Sp(4, \mathbb{R})$ at a 5-th root of unity, see [36]). We note here that the standard invariants $\mathcal{P} = (P_k)_{k\geq 0}$ of *all* BMW subfactors are generated by a single non-trivial operator in P_2 and that they satisfy the condition $\dim P_3 \leq 15$.

The proof of this theorem uses in a crucial way theorem 5.1 and the tight restrictions imposed by compatibility of the rotation of period 3 on P_3 and the algebra structure.

The next phase of this enumeration project will be to tackle the case $d = 14$. Here we know that the quantum $Sp(4, \mathbb{R})$ specialization of the BMW algebra will give examples with a free parameter. We do expect however, that the general ideas of [9] and [10] will enable us to enumerate all such subfactor planar algebras.

References

[1] M. Asaeda & U. Haagerup, *Exotic subfactors of finite depth with Jones indices* $(5 + \sqrt{13})/2$ *and* $(5 + \sqrt{17})/2$, Comm. Math. Phys. **202** (1999), 1–63.

[2] T. Banica, Representations of compact quantum groups and subfactors, *J. Reine Angew. Math.* 509 (1999), 167–198.

[3] D. Bisch, A note on intermediate subfactors, *Pacific Journal of Math.* 163 (1994), 201-216.

[4] D. Bisch, Bimodules, higher relative commutants and the fusion algebra associated to a subfactor, *The Fields Institute for Research in Math. Sciences Commun. Series*, vol. 13, AMS, Providence, Rhode Island, 1997, 13-63.

[5] D. Bisch, An example of an irreducible subfactor of the hyperfinite II_1 factor with rational, noninteger index, *J. Reine Angew. Math.* 455 (1994), 21-34.

[6] D. Bisch & U. Haagerup, Composition of subfactors: new examples of infinite depth subfactors, *Ann. scient. Éc. Norm. Sup.* 29 (1996), 329-383.

[7] D. Bisch & V.F.R. Jones, Algebras associated to intermediate subfactors, *Invent. Math.* 128 (1997), 89-157.

[8] D. Bisch & V.F.R. Jones, A note on free composition of subfactors, *"Geometry and Physics"*, vol. 184, Marcel Dekker, Lecture Notes in Pure and Applied Mathematics, 1997, 339-361.

[9] D. Bisch & V.F.R. Jones, Singly generated planar algebras of small dimension, *Duke Math. Journal* 101 (2000), 41-75.

[10] D. Bisch & V.F.R. Jones, Singly generated planar algebras of small dimension, Part II, *Advances in Math.* (to appear).

[11] D. Bisch & S. Popa, Examples of subfactors with property T standard invariant, *Geom. Funct. Anal.* 9 (1999), 215-225.

[12] D. Bisch & S. Popa, A continuous family of non-isomorphic irreducible hyperfinite subfactors with the same standard invariant, *in preparation..*

[13] A. Connes, *Noncommutative geometry*, Academic Press, 1994.

[14] D. Evans & Y. Kawahigashi, *Quantum symmetries on operator algebras*, Oxford University Press, 1998.

[15] P. Di Francesco, New integrable lattice models from Fuss-Catalan algebras, *Nuclear Phys. B* 532 (1998), 609-634.

[16] F. Goodman & P. de la Harpe & V.F.R. Jones, *Coxeter graphs and towers of algebras*, Springer Verlag, MSRI publications, 1989.

[17] U. Haagerup, Principal graphs of subfactors in the index range $4 < [M : N] < 3 + \sqrt{2}$, Subfactors (Kyuzeso, 1993), World Sci. Publishing, River Edge, NJ, 1994, 1–38.

[18] M. Izumi, Applications of fusion rules to classification of subfactors, *Publ. RIMS, Kyoto Univ.* 27 (1991), 953-994.

[19] V.F.R. Jones, Index for subfactors, *Invent. Math.* 72 (1983), 1-25.

[20] V.F.R. Jones, Hecke algebra representations of braid groups and link polynomials, *Ann. of Math.* 126, 335-388.

[21] V.F.R. Jones, On a certain value of the Kauffman polynomial, *Comm. Math. Phys.* 125 (1989), 459–467.

[22] V.F.R. Jones, Planar algebras I, *preprint.*

[23] V.F.R. Jones, The planar algebra of a bipartite graph, *Knots in Hellas '98 (Delphi)*, World Sci. Publishing, 2000, 94-117.

[24] V.F.R. Jones, The annular structure of subfactors, *Enseign. Math.* (to appear).

[25] L. Kauffman, State models and the Jones polynomial, *Topology* 26 (1987), 395-407.

[26] Z. Landau, Fuss-Catalan algebras and chains of intermediate subfactors, *Pacific J. Math.* 197 (2001), 325-36.

[27] Z. Landau, Exchange relation planar algebras, *preprint* (2000).

[28] J.P. May, Definitions: operads, algebras and modules, *Contemporary Mathematics* 202 (1997), 1-7.

[29] M. J. Martins & B. Nienhuis, Applications of Temperley-Lieb algebras to Lorentz lattice gases, *J. Phys. A* 31 (1998), L723–L729.

[30] A. Ocneanu, Quantized group string algebras and Galois theory for operator algebras, in Operator Algebras and Applications 2, *London Math. Soc. Lect. Notes Series* 136 (1988), 119-172.

[31] M. Pimsner & S. Popa, Entropy and index for subfactors, *Ann. scient. Ec. Norm. Sup.* 19 (1986), 57-106.

[32] S. Popa, Classification of subfactors: reduction to commuting squares, *Invent. Math.* 101 (1990), 19-43.

[33] S. Popa, Classification of amenable subfactors of type II, *Acta Math.* 172 (1994), 352-445.

[34] S. Popa, An axiomatizaton of the lattice of higher relative commutants, *Invent. Math.* 120 (1995), 427-445.

[35] A. Wassermann, Operator algebras and conformal field theory III, *Invent. Math.* 92 (1998), 467-538.

[36] H. Wenzl, Quantum groups and subfactors of type B, C and D, *Comm. Math. Phys* 133, 383-432.

[37] H. Wenzl, C^* tensor categories from quantum groups, *J. Amer. Math. Soc.* 11 (1998), 261-282.

[38] F. Xu, Standard λ-lattices from quantum groups, *Invent. Math.* 134 (1998), 455–487.

ICM 2002 · Vol. II · 787–794

Free Probability, Free Entropy and Applications to von Neumann Algebras

Liming Ge*

This talk is organized as follows: First we explain some basic concepts in non-commutative probability theory in the frame of operator algebras. In Section 2, we discuss related topics in von Neumann algebras. Sections 3 and 4 contain some of the key ideas and results in free probability theory. Last section states some of the important applications of free probability theory.

1. Non-commutative probability spaces

In general, a non-commutative probability space is a pair $(\mathcal{A}, \tau)$, where $\mathcal{A}$ is a unital algebra (over the field of complex numbers $\mathbf{C}$) and τ a linear functional with $\tau(I) = 1$, where I is the identity of $\mathcal{A}$. Elements of $\mathcal{A}$ are called *random variables.* Since positivity is a key concept in (classical) probability theory, this can be captured by assuming that $\mathcal{A}$ is a * algebra and τ is positive (i.e., a state). Elements of the form A^*A are called *positive* (random variables).

A state τ is a *trace* if $\tau(AB) = \tau(BA)$. We often require that τ be a faithful trace (τ corresponds to the classical probability measure, or the integral given by the measure). In this talk, we always assume that $\mathcal{A}$ is a unital * algebra over $\mathbf{C}$ and τ a faithful state on $\mathcal{A}$. Subalgebras of $\mathcal{A}$ are always assumed unital * subalgebras.

Examples of noncommutative probability spaces often come from operator algebras on a Hilbert space and the states used here are usually vector states.

A *C*-probability space* is a pair $(\mathcal{A}, \tau)$, where $\mathcal{A}$ is a unital C*-algebra (norm closed subalgebra of $\mathcal{B}(\mathcal{H})$) and τ is a state on $\mathcal{A}$.

A *W*-probability space* is a pair $(\mathcal{M}, \tau)$ consisting of a von Neumann algebra $\mathcal{M}$ (strong-operator closed C*-subalgebra of $\mathcal{B}(\mathcal{H})$) and a *normal* (i.e., countably additive) state τ on $\mathcal{M}$.

The following are some more basic concepts:

Independence: In a noncommutative probability space $(\mathcal{A}, \tau)$, a family $\{\mathcal{A}_j\}$ of subalgebras $\mathcal{A}_j$ of $\mathcal{A}$ is *independent* if the subalgebras commute with each other and, for $n \in \mathbf{N}$, $\tau(A_1 \cdots A_n) = \tau(A_1) \cdots \tau(A_n)$ for all A_k in $\mathcal{A}_{j_k}$ and $j_k \neq j_l$ whenever $k \neq l$.

* Academy of Mathematics and System Science, CAS, Beijing 100080, China. Department of Mathematics UNH, Durham, NH 03824, USA. E-mail: liming@math.unh.edu

This independence gives a "tensor-product" relation among subalgebras $\mathcal{A}_j$: if $\mathcal{A}$ is generated by $\mathcal{A}_j$, then $\mathcal{A} \cong \otimes_j \mathcal{A}_j$ (in the case of C*- or W*-probability spaces, the tensor-product shall reflect the corresponding topological structures on $\mathcal{A}$ and $\mathcal{A}_j$).

Distributions and moments: Given $(\mathcal{A}, \tau)$, for A in $\mathcal{A}$, we define a map $\mu_A : \mathbf{C}[x] \to \mathbf{C}$ by $\mu_A(p(x)) = \tau(p(A))$. Then μ_A is the *distribution* of A. For $A_1, \ldots, A_n$ in $\mathcal{A}$, the *joint distribution* $\mu_{A_1,\ldots,A_n} : \mathbf{C}\langle x_1, \ldots, x_n\rangle \to \mathbf{C}$ is given by

$$\mu_{A_1,\ldots,A_n}(p(x_1, \ldots, x_n)) = \tau(p(A_1, \ldots, A_n)).$$

If p is a monomial, $\tau(p(A_1, \ldots, A_n))$ is called a *(p-)moment.* When random variables are non self-adjoint, one also considers (joint) * distributions of random variables, that can be defined in a similar way. In this case, there is a natural identification of $\mathbf{C}\langle x_1, \ldots, x_n, x_1^*, \ldots, x_n^*\rangle$ with the semigroup algebra $\mathbf{C}S_{2n}$, where S_{2n} is the free semigroup on $2n$ generators. Monomials are given by words in S_{2n}.

Conditional Expectations: Suppose $\mathcal{B}$ is a subalgebra of $\mathcal{A}$. A *conditional expectation* from $\mathcal{A}$ onto $\mathcal{B}$ is a $\mathcal{B}$-bimodule map (a projection of norm one in the case of C*-algebras) of $\mathcal{A}$ onto $\mathcal{B}$ so that the restriction on $\mathcal{B}$ is the identity map.

Many other concepts in probability theory and measure theory can be generalized to operator algebras, especially von Neumann algebras which can be regarded as non-commutative measure spaces. For basic operator algebra theory, we refer to [KR] and [T].

2. GNS representation and von Neumann algebras

Given a C*-probability space $(\mathcal{A}, \tau)$, one defines an inner product $\langle A, B\rangle = \tau(B^*A)$ on $\mathcal{A}$. Let $L^2(\mathcal{A}, \tau)$ be the Hilbert space obtained by the completion of $\mathcal{A}$ under the L^2-norm given by this inner product. Then $\mathcal{A}$ acts on $L^2(\mathcal{A}, \tau)$ by left multiplication. This representation of $\mathcal{A}$ on the Hilbert space $L^2(\mathcal{A}, \tau)$ is called the *GNS representation.* In a similar way, one can define $L^p(\mathcal{A}, \tau)$, where $\|A\|_p = \tau(|A|^p)^{1/p} = \tau((A^*A)^{p/2})^{1/p}$. The von Neumann algebra generated by $\mathcal{A}$ (or the strong-operator closure of $\mathcal{A}$) is sometimes denoted by $L^\infty(\mathcal{A}, \tau)$ ($\subset L^p(\mathcal{A}, \tau)$, $p \geq 1$). All von Neumann algebras admit such a form. Any von Neumann algebra is a (possibly, continuous) direct sum of "simple" algebras, or factors (algebras with a trivial center). Von Neumann algebras that admit a faithful (finite) trace are said to be *finite*. The classification of (infinite-dimensional) finite factors has become the central problem in von Neumann algebras.

Murray and von Neumann [MN] also separate factors into three types:

Type I: Factors contain a minimal projection. They are isomorphic to full matrix algebras $M_n(\mathbf{C})$ or $\mathcal{B}(\mathcal{H})$.

Type II: Factors contain a "finite" projection but without minimal projections: it is said to be of type II_1 when the identity I is a finite projection; of type II_∞ when I is infinite. Every type II_∞ is the tensor product of $\mathcal{B}(\mathcal{H})$ with a factor of type II_1.

Type III: Every (non zero) projection is infinite.

Examples of von Neumann algebras: 1) Let $\mathcal{A} = \mathbf{C}G$ for some discrete group G, $\mathcal{H}$ be $l^2(G)$ and τ be the vector state given by the vector that takes value 1 at g and 0 elsewhere. Then τ is a trace, $l^2(G) = L^2(\mathcal{A}, \tau)$ and the weak (or strong) operator closure of $\mathcal{A}$ is called the group von Neumann algebra, denoted by $\mathcal{L}_G$. We have that $\mathcal{L}_G$ is a factor if and only if each conjugacy class of G other than the identity is infinite (i.c.c.). For example, the free group F_n ($n \geq 2$, on n generators) is such an i.c.c. group.

2) Suppose (Ω, μ) is a measure space with a σ-finite measure μ, G is a group and α is a measurability preserving action of G on Ω. Formally, we have an algebra $L^\infty(\Omega, \mu)G$ $(= \mathcal{A})$ similar to the group algebra definition: $(\varphi(x)g)(\psi(x)h) = \varphi(x)\psi(g^{-1}(x))gh$, for $\varphi, \psi \in L^\infty(\Omega, \mu)$ and $g, h \in G$. Assume that G acts freely (i.e., for any g in G with $g \neq e$, the set $\{x \in \Omega : g(x) = x\}$ has measure zero). Define an action of $\mathcal{A}$ on the Hilbert space $\bigoplus_{g \in G} L^2(\Omega, \mu)g$ by left multiplication (which is induced by the multiplication in $\mathcal{A}$), where $L^2(\Omega, \mu)g$ is an isomorphic copy of $L^2(\Omega, \mu)$. Then the von Neumann algebra generated by $\mathcal{A}$ is called the cross product von Neumann algebra, denoted by $L^\infty(\Omega, \mu) \times_\alpha G$. This cross product is a factor if and only if α is ergodic. If α is ergodic and *not* a measure preserving action, then $L^\infty(\Omega, \mu) \times_\alpha G$ is a factor of type III. Type II factors are obtained from measure preserving actions (with Ω a non atomic measure space) and the finiteness of μ gives rise to type II_1 factors.

The above 1) and 2) are the two basic constructions of von Neumann algebras (given by Murray and von Neumann [MN]). A. Connes [C] shows that there are finite factors that cannot be constructed by 1). It was a longstanding open problem whether every (finite) von Neumann algebra can be obtained by using the construction in 2). Using free probability theory, especially the notion of free entropy, Voiculescu [V2] gives a negative answer to this question. We shall discuss some details later in the talk.

In recent years, the focus of studies of von Neumann algebras is centered on factors of type II_1. Many of the unsolved problems in operator algebras are also reduced to this class. We end this section with two of the (still) open problems from the list of 20 questions asked by Kadison in 1967.

1. The weak-operator closure of the left regular representation of the free (non-abelian) group on two or more generators is a factor of type II_1. Are these factors isomorphic for different numbers of generators?

2. Is each factor generated by two self-adjoint operators? —each von Neumann algebra? —the factor arising from the free group on three generators? —is each von Neumann algebra finitely generated?

Three of those 20 problems were answered in the last ten years by using free probability and free entropy. We explain some of the theory involved in the following two sections.

3. Free independence

Suppose $(\mathcal{A}, \tau)$ is a C*-probability space. We assume that τ is a trace. A family $\mathcal{A}_\iota$, $\iota \in \mathbf{I}$, of unital * subalgebras of $\mathcal{A}$ are called *free* with respect to τ if $\tau(A_1 A_2 \cdots A_n) = 0$ whenever $A_j \in \mathcal{A}_{\iota_j}$, $\iota_1 \neq \cdots \neq \iota_n$ (ι_1 and ι_3 may be the same) and $\tau(A_j) = 0$ for $1 \leq j \leq n$ and every n in $\mathbf{N}$. A family of subsets (or elements) of $\mathcal{A}$ are said to be *free* if the unital * subalgebras they generate are free.

Note that freeness is a highly noncommutative notion, the non-commutativity (or algebraic freeness) of free random variables is encoded in the definition. Recall some basic concepts in free probability theory.

Semicircular elements: The Gaussian laws in classical theory is replaced by the semicircular laws. The *semicircular law* centered at a and of radius r is the distribution $\mu_{a,r} : \mathbf{C}[x] \to \mathbf{C}$ such that

$$\mu_{a,r}(\varphi(x)) = \frac{2}{\pi r^2} \int_{a-r}^{a+r} \varphi(t)\sqrt{r^2 - (t-a)^2}dt,$$

for each $\varphi(x)$ in $\mathbf{C}[x]$. A self-adjoint random variable A in $(\mathcal{A}, \tau)$ is said to be *(standard) semicircular* if its distribution is $\mu_{0,1}$. An element $X = A + iB$ is *circular* if A and B are free semicircular. The following theorem is proved by D. Voiculescu (see [VDN]).

Free Central Limit Theorem: *Let $\{A_j\}_{j=1}^{\infty}$ be a free family of identically distributed random variables in $(\mathcal{A}, \tau)$ with $\tau(A_j) = 0$ and $\tau(A_j^2) = \frac{r^2}{4}$ for some positive number r. Then the distribution of $\frac{A_1 + \cdots + A_n}{\sqrt{n}}$ converges to the semicircular law $\mu_{0,r}$.*

Classical independence corresponds to tensor products; while the above free independence introduced by Voiculescu is given by certain free products. Recall some examples of such freeness.

Free products: If $G = G_1 * G_2$, then $\mathcal{L}_{G_1}$ and $\mathcal{L}_{G_2}$ are free in $\mathcal{L}_G$, here the trace is the one given by the unit vector associated with any group element.

Let $(\mathcal{A}_1, \tau_1)$ and $(\mathcal{A}_2, \tau_2)$ be two W*-probability spaces. Suppose $\mathcal{A}_0$ is the (amalgamated algebraic) free product of $\mathcal{A}_1$ and $\mathcal{A}_2$ (over the scalars). Then there is a unique τ on $\mathcal{A}_0$ such that $\mathcal{A}_1$ and $\mathcal{A}_2$ are free with respect to τ and the restrictions of τ on $\mathcal{A}_1$ and $\mathcal{A}_2$ equal to τ_1 and τ_2, respectively. Let $\mathcal{A}$ be the weak operator closure of $\mathcal{A}_0$ acting on $L^2(\mathcal{A}_0, \tau)$. Then $\mathcal{A}$ is called the (reduced von Neumann algebra) free product of $\mathcal{A}_1$ and $\mathcal{A}_2$ (with respect to τ), denoted by $\mathcal{A}_1 * \mathcal{A}_2$ (and $\tau = \tau_1 * \tau_2$). For example, $\mathcal{L}_{F_2} \cong L^\infty[0,1] * L^\infty[0,1]$.

Full Fock space construction: Let $\mathcal{H}_0$ be a real Hilbert space and $\mathcal{H}$ be $\mathcal{H}_0 \otimes \mathbf{C}$. Its *full Fock space* is

$$\mathcal{T}(\mathcal{H}) = \mathbf{C}1 \oplus \bigoplus_{n \geq 1} \mathcal{H}^{\otimes n}.$$

For $h \in \mathcal{H}_0$, let the *left creation operator* $l(h) \in \mathcal{B}(\mathcal{T}(\mathcal{H}))$ be given by $l(h)\xi = h \otimes \xi$. Then $l(h)^*1 = 0$ and $l(h)^*\xi_1 \otimes \xi_2 \otimes \cdots \otimes \xi_n = \langle \xi_1, h\rangle \xi_2 \otimes \cdots \otimes \xi_n$, so $l(h_1)^* l(h_2) = \langle h_2, h_1\rangle I$. Let $C^*(l(\mathcal{H}_0))$ (or $W^*(l(\mathcal{H}_0))$) be the C*- (or W*-) algebra generated by $\{l(h) | h \in \mathcal{H}_0\}$. Let $\omega_{\mathcal{H}}$ be the vector state given by the vector $1 \in \mathcal{T}(\mathcal{H})$. Here 1 is

called the *vacuum vector* and $\omega_{\mathcal{H}}$ the *vacuum state.* If $\mathcal{H}_1$ and $\mathcal{H}_2$ are orthogonal subspaces of $\mathcal{H}_0$, then $C^*(l(\mathcal{H}_1))$ and $C^*(l(\mathcal{H}_2))$ are free (with respect to $\omega_{\mathcal{H}}$). If h is a unit vector in $\mathcal{H}_0$, then $(l(h)+l(h)^*)/2$ is semicircular with distribution $\mu_{0,1}$.

Gaussian Random Matrices: Let $X(s,n)=(f_{ij}(s,n))$ in $M_n(L^\infty[0,1])$ be real random matrices, where $n\in\mathbf{N}$ and $s\in S$ for some index set S. Assume that $f_{ij}(s,n)=f_{ji}(s,n)$ and $\{f_{ij}(s,n):i,j,s\}$ (given each n) is a family of independent Gaussian $(0,1/n)$ random variables. Let D_n be a constant diagonal matrix in $M_n(\mathbf{R})$ having a limit distribution (as $n\to\infty$). Then $\{X(s,n)\}\cup\{D_n\}$ is asymptotically free as $n\to\infty$ and $\{X(s,n):s\in S\}$ converges in distribution to a free semicircular family. As a corollary, Voiculescu shows that $\mathcal{L}_{F_5}\otimes M_2(\mathbf{C})\cong\mathcal{L}_{F_2}$. Moreover $\mathcal{L}_{F_r}\otimes M_n(\mathbf{C})\cong\mathcal{L}_{F_{1+\frac{r-1}{n^2}}}$, for any real number r, and $\mathcal{L}_{F_r}*\mathcal{L}_{F_s}\cong\mathcal{L}_{F_{r+s}}$. Now we know that either $\mathcal{L}_{F_r}$, $r>1$, are all isomorphic to each other or they are all non isomorphic factors (see [D] and [R]).

Further studies of free probability theory have been pursued by many people in several directions, such as infinitely divisible laws, free brownian motion, etc. (we refer to [B], [VDN] and [HP] for details).

4. Free entropy

Free entropy is a non commutative analogue of classical Shannon entropy. First we recall the definition of entropy and its basic properties.

Classical Entropy: Let (Ω,Σ,μ) be a probability space with probability measure μ and $f_1,\dots,f_n:\Omega\to\mathbf{R}$ be random variables. Suppose φ is the density function on $\mathbf{R}^n$ corresponding to the joint distribution of $f_1,\dots,f_n$. Then the entropy:

$$H(f_1,\dots,f_n)=-\int_{\mathbf{R}^n}\varphi(t_1,\dots,t_n)\log\varphi(t_1,\dots,t_n)dt_1\cdots dt_n.$$

Here are two important properties of entropy: $H(f_1,\dots,f_n)=H(f_1)+\cdots+H(f_n)$ if and only if $f_1,\dots,f_n$ are independent; when assuming that $E(f_j^2)=1$, H is maximal if and only if $f_1,\dots,f_n$ are Gaussian independent $(0,1)$ random variables.

Free entropy: Let $X_1,\dots,X_n$ be self-adjoint random variables in $(\mathcal{A},\tau)$. For any $\varepsilon>0$, when k large, there may be self-adjoint matrices $A_1,\dots,A_n$ in $M_k(\mathbf{C})$ such that *"the algebra generated by X_j's looks like the algebra generated by A_j's within ε."* More precisely, for any $\varepsilon>0$, large $m\in\mathbf{N}$ and any monomial p in $\mathbf{C}\langle x_1,\dots,x_n\rangle$ with degree less than or equal to m, choose k large enough so that

$$(*)\qquad |\tau(p(X_1,\dots,X_n))-\tau_k(p(A_1,\dots,A_n))|<\varepsilon.$$

Let $\Gamma_R(X_1,\dots,X_n;m,k,\varepsilon)$ be the set of all self-adjoint matrices $(A_1,\dots,A_n)$ in $M_k(\mathbf{C})$, with $\|A_j\|\le R$, such that $(*)$ holds. The limit of the "normalized" measurement of $\Gamma_R(X_1,\dots,X_n;m,k,\varepsilon)$ is called the free entropy of $X_1,\dots,X_n$. Voiculescu [V1] shows that this limit is independent of R when it is larger than $\max\{\|X_j\|:j=1,\dots,n\}$. Here we will fix such a constant R and use $\Gamma(X_1,\dots,X_n;$

$m, k, \varepsilon)$ to denote $\Gamma_R(X_1, \ldots, X_n; m, k, \varepsilon)$. Let *vol* be the euclidean volume in real euclidean space $(M_k(\mathbf{C})^{s.a.})^n$ (here "s.a." denote the self-adjoint part and the euclidean norm $\|A\|_e^2 = \mathrm{Tr}(A^2)$). Now we define, successively,

$$\chi(X_1, \ldots, X_n; m, k, \varepsilon) = \log vol(\Gamma(X_1, \ldots, X_n; m, k, \varepsilon)),$$
$$\chi(X_1, \ldots, X_n; m, \varepsilon) = \limsup_{k \to \infty}(k^{-2}\chi(X_1, \ldots, X_n; m, k, \varepsilon) + \frac{n}{2}\log k),$$
$$\chi(X_1, \ldots, X_n) = \inf\{\chi(X_1, \ldots, X_n; m, \varepsilon) : m \in \mathbf{N}, \varepsilon > 0\}.$$

We call $\chi(X_1, \ldots, X_n)$ the *free entropy* of $(X_1, \ldots, X_n)$.

The following are some basic properties of free entropy (proved in [V1]):

(i) $\chi(X_1, \ldots, X_n) \leq \frac{n}{2}\log(2\pi e C^2 n^{-1})$;

(ii) $\chi(X_1) = \iint \log|s-t| d\mu_1(s) d\mu_1(t) + \frac{3}{4} + \frac{1}{2}\log 2\pi$, where μ_1 is the (measure on the spectrum of X_1 corresponding to the) distribution of X_1;

(iii) $\chi(X_1, \ldots, X_n) = \chi(X_1) + \cdots + \chi(X_n)$ when $X_1, \ldots, X_n$ are free random variables.

Voiculescu also introduces a notion of free entropy dimension $\delta(X_1, \ldots, X_n)$ which is given by

$$\delta(X_1, \ldots, X_n) = n + \limsup_{\varepsilon \to 0} \frac{\chi(X_1 + \varepsilon S_1, \ldots, X_n + \varepsilon S_n : S_1, \ldots, S_n)}{|\log \varepsilon|}$$

where $\{S_1, \ldots, S_n\}$ is a standard semicircular family and $\{X_1, \ldots, X_n\}$ and $\{S_1, \ldots, S_n\}$ are free.

5. Applications

In this section, we review some of the applications of free entropy in von Neumann algebras.

Theorem 1. ([V1]) *Let $S_1, \ldots, S_n$ be a standard (0,1) free semicircular family. Then*

$$\chi(S_1, \ldots, S_n) = n(\frac{1}{2} + \log\frac{\sqrt{2\pi}}{2});$$
$$\delta(S_1, \ldots, S_n) = n.$$

Note that $S_1, \ldots, S_n$ generate $\mathcal{L}_{F_n}$ as a von Neumann algebra.

Theorem 2. ([V2]) *If a finite von Neumann algebra $\mathcal{M}$ (with a trace) has a Cartan subalgebra, then, for any self-adjoint generators $X_1, \ldots, X_n$ of $\mathcal{M}$, $\delta(X_1, \ldots, X_n) \leq 1$.*

This implies that free group factors do not contain any Cartan subalgebras which answers a longstanding question in the subject.

The following result generalizes Voiculescu's result and answers Problem 11 on Kadison's 1967 problem list (unpublished).

Theorem 3. ([G1]) *If a finite von Neumann algebra $\mathcal{M}$ (with a trace) has a simple maximal abelian subalgebra, then, for any self-adjoint generators $X_1, \ldots, X_n$ of $\mathcal{M}$, $\delta(X_1, \ldots, X_n) \leq 2$.*

Furthermore, we prove the following result which shows the existence of a separable prime factor (the one that is not the tensor product of two factors of the same type).

Theorem 4. ([G2]) *If $\mathcal{M} = \mathcal{M}_1 \bar{\otimes} \mathcal{M}_2$ for some infinite dimensional finite von Neumann algebras $\mathcal{M}_1$ and $\mathcal{M}_2$, then, for any self-adjoint generators $X_1, \ldots, X_n$ of $\mathcal{M}$, $\delta(X_1, \ldots, X_n) \leq 1$.*

It is an outstanding open question whether the free entropy dimension is an invariant for von Neumann algebras. Following Voiculescu's result in [V3], we prove the following:

Theorem 5. ([GS]) *For any self-adjoint generators $X_1, \ldots, X_n$ of $\mathcal{L}_{SL_m(\mathbf{Z})}$, $m \geq 3$, $\delta(X_1, \ldots, X_n) \leq 1$.*

The above theorem is not true for $SL_2(\mathbf{Z})$.

References

[B] P. Biane, Proceedings of ICM2002.

[C] A. Connes, *Sur la classification des facteurs de type* II_1, (French. English summary) C. R. Acad. Sci. Paris Sér. A-B **281** (1975), A13–A15.

[D] K. Dykema, *Free products of hyperfinite von Neumann algebras and free dimension,* Duke Math. J., **69** (1993), 97–119.

[G1] L. Ge, *Applications of free entropy to finite von Neumann algebras,* Amer. J. Math., **119** (1997), 467–485.

[G2] L. Ge, *Applications of free entropy to finite von Neumann algebras, II,* Ann. of Math., **147** (1998), 143–157.

[GS] L. Ge and J. Shen, *Free entropy and property T factors*, Proc. Nat. Acad. Sci. (USA), **97** (2000), 9881–9885.

[HP] F. Hiai and D. Petz, *"The Semicircle Law, Free Random Variables and Entropy"*, Mathematical Surveys and Monographs, **77**, American Mathematical Society, Providence, RI, 2000.

[KR] R. Kadison and J. Ringrose, *"Fundamentals of the Operator Algebras,"* vols. I and II, Academic Press, Orlando, 1983 and 1986.

[MN] F. J. Murray and J. von Neumann, *On rings of operators*, Ann. of Math., **37** (1936), 116–229.

[R] F. Radulescu, *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.

[T] M. Takesaki, *"Theory of Operator Algebras*, vol I, Springer-Verlag, New York-Heidelberg, 1979.

[V1] D. Voiculescu, *The analogues of entropy and of Fisher's information measure in free probability theory II,* Invent. Math., **118** (1994), 411-440.

[V2] D. Voiculescu, *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.

[V3] D. Voiculescu, *Free entropy dimension* ≤ 1 *for some generators of property T factors of type* II_1, J. Reine Angew. Math. **514** (1999), 113–118.

[VDN] D. Voiculescu, K. Dykema and A. Nica, *"Free Random Variables,"* CRM Monograph Series, vol. 1, AMS, Providence, R.I., 1992.

ICM 2002 · Vol. II · 795–812

Banach KK-theory and the Baum-Connes Conjecture

V. Lafforgue*

Abstract

The report below describes the applications of Banach KK-theory to a conjecture of P. Baum and A. Connes about the K-theory of group C^*-algebras, and a new proof of the classification by Harish-Chandra, the construction by Parthasarathy and the exhaustion by Atiyah and Schmid of the discrete series representations of connected semi-simple Lie groups.

2000 Mathematics Subject Classification: 19K35, 22E45, 46L80.
Keywords and Phrases: Kasparov's KK-theory, Baum-Connes conjecture, Discrete series.

This report is intended to be very elementary. In the first part we outline the main results in Banach KK-theory and the applications to the Baum-Connes conjecture. In the second part we show how the Baum-Connes conjecture for connected semi-simple Lie groups can be applied to recover the classification of the discrete series representations.

1. Banach KK-theory and the Baum-Connes conjecture

There are many surveys on Kasparov's KK-theory and the Baum-Connes conjecture (see [4, 48, 49, 21, 27, 13, 54]) and on Banach KK-theory ([49, 38]).

1.1. Generalized Fredholm modules

We wish to define A-linear Fredholm operators (where A is a Banach algebra), with an index in $K_0(A)$. If $A = \mathbb{C}$, this index should be the usual index of $\mathbb{C}$-linear Fredholm operators in $K_0(\mathbb{C}) = \mathbb{Z}$.

*Institut de Mathématiques de Jussieu, 175 rue du Chevaleret, 75013 Paris, France. E-mail: vlafforg@math.jussieu.fr

We define a Banach algebra as a (non necessarily unital) $\mathbb{C}$-algebra A that is complete for a norm $\|.\|$ satisfying $\|ab\| \leq \|a\|\|b\|$ for any $a, b \in A$. If A and B are Banach algebras a morphism $\theta : A \to B$ is an algebra morphism such that $\|\theta(a)\| \leq \|a\|$ for any $a \in A$.

K_0 and K_1 are two covariant functors from the category of Banach algebras to the category of abelian groups. If X is a locally compact space and $C_0(X)$ the algebra of continuous functions vanishing at infinity, $K_0(C_0(X))$ and $K_1(C_0(X))$ are the Atiyah-Hirzebruch K-theory groups. For technical reasons we shall restrict ourselves to unital Banach algebras in this subsection.

Let A be a unital Banach algebra.

A right A-module E is finitely generated projective if and only if it is a direct summand in A^n for some integer n. The set of isomorphism classes of right finitely generated projective A-modules is a semigroup because the direct sum of two right finitely generated projective A-modules is a right finitely generated projective A-module. Then $K_0(A)$ is the universal group associated to this semigroup (*i.e.* the group of formal differences of elements of the semigroup). If $\theta : A \to B$ is a morphism of unital Banach algebras, and E is a right finitely generated projective A-module then $E \otimes_A B$ is a right finitely generated projective B-module and this defines $\theta_* : K_0(A) \to K_0(B)$.

There is another definition of $K_0(A)$ for which the functoriality is even more obvious : $K_0(A)$ is the quotient of the free abelian group generated by all idempotents p in $M_k(A)$ for some integer k, by the relations $\left[\begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} \right] = [p] + [q]$ for any idempotents $p \in M_k(A)$ and $q \in M_l(A)$ and $[p] = [q]$ if p, q are idempotents of $M_k(A)$ and are connected by a path of idempotents in $M_k(A)$ and $[0] = 0$ where 0 is the idempotent 0 in $M_k(A)$. The link with the former definition is that any idempotent $p \in M_k(A)$ acts on the left on A^k as a projector P and $\mathrm{Im}P$ is a right finitely generated projective A-module (it is a direct summand in the right A-module A^k).

The following construction was performed for C^*-algebras by Mischenko and Kasparov, in connection with the Novikov conjecture ([43, 28]). We adapt it to Banach algebras.

A right Banach A-module is a Banach space (with a given norm $\|.\|_E$) equipped with a right action of A such that $1 \in A$ acts by identity and $\|xa\|_E \leq \|x\|_E\|a\|_A$ for any $x \in E$ and $a \in A$. Let E and F be right Banach A-modules. A morphism $u : E \to F$ of right Banach A-modules is a continuous $\mathbb{C}$-linear map such that $u(xa) = u(x)a$ for any $x \in E$ and $a \in A$. The space $\mathcal{L}_A(E, F)$ of such morphisms is a Banach space with norm $\|u\| = \sup_{x \in E, \|x\|_E = 1} \|u(x)\|_F$. A morphism $u \in \mathcal{L}_A(E, F)$ is said to be "A-rank one" if $u = w \circ v$ with $v \in \mathcal{L}_A(E, A)$ and $w \in \mathcal{L}_A(A, F)$. The space $\mathcal{K}_A(E, F)$ of A-compact morphisms is the closed vector span of A-rank one morphisms in $\mathcal{L}_A(E, F)$. If $E = F$, $\mathcal{L}_A(E) = \mathcal{L}_A(E, E)$ is a Banach algebra and $K_A(E) = K_A(E, E)$ is a closed ideal in it.

Definition 1.1.1 *A Fredholm module over A is the data of a $\mathbb{Z}/2$ graded right Banach A-module E and an odd morphism $T \in \mathcal{L}_A(E)$ such that $T^2 - \mathrm{Id}_E \in \mathcal{K}_A(E)$.*

In other words $E = E_0 \oplus E_1$, $T = \begin{pmatrix} 0 & v \\ u & 0 \end{pmatrix}$ and $u \in \mathcal{L}_A(E_0, E_1)$ and $v \in$

$\mathcal{L}_A(E_1, E_0)$ satisfy $vu - \mathrm{Id}_{E_0} \in \mathcal{K}_A(E_0)$ and $uv - \mathrm{Id}_{E_1} \in \mathcal{K}_A(E_1)$.

If (E, T) is a Fredholm module over A and $\theta : A \to B$ a unital morphism then $(E \otimes_A B, T \otimes 1)$ is a Fredholm module over B (here $E \otimes_A B$ is the completion of $E \otimes_A^{\mathrm{alg}} B$ for the maximal Banach norm such that $\|x \otimes b\| \leq \|x\|_E \|b\|_B$ for $x \in E$ and $b \in B$).

Let $A[0,1]$ be the Banach algebra of continuous functions from $[0,1]$ to A with the norm $\|f\| = \sup_{t \in [0,1]} \|f(t)\|_A$ and $\theta_0, \theta_1 : A[0,1] \to A$ the evaluations at 0 and 1. Two Fredholm modules on A are said to be homotopic if they are the images by θ_0 and θ_1 of a Fredholm module over $A[0,1]$.

Theorem 1.1.2 *There is a functorial bijection between $K_0(A)$ and the set of homotopy classes of Fredholm modules over A, for any unital Banach algebra A.*

Let (E_0, E_1, u, v) be a Fredholm module over A. Its index, *i.e.* the corresponding element in $K_0(A)$, is constructed as follows. It is possible to find $n \in \mathbb{N}$ and $w \in \mathcal{K}_A(A^n, E_1)$ such that $(u, w) \in \mathcal{L}_A(E_0 \oplus A^n, E_1)$ is surjective. Its kernel is then finitely generated projective and the index is the formal difference of $\mathrm{Ker}((u, w))$ and A^n.

An ungraded Fredholm module over A is the data of a (ungraded) right Banach module E over A, and $T \in \mathcal{L}_A(E)$ such that $T^2 - \mathrm{Id}_E \in \mathcal{K}_A(E)$. There is a functorial bijection between $K_1(A)$ and the set of homotopy classes of ungraded Fredholm modules.

For a non-unital algebra A, $K_0(A) = \mathrm{Ker}(K_0(\tilde{A}) \to K_0(\mathbb{C}) = \mathbb{Z})$ and $K_1(A) = K_1(\tilde{A})$ where $\tilde{A} = A \oplus \mathbb{C}1$. In particular every idempotent in $M_k(A)$ gives a class in $K_0(A)$ but in general not all classes in $K_0(A)$ are obtained in this way. The definition of a Fredholm module should be slightly modified for non-unital Banach algebras, but the theorem 1.1.2 remains true.

1.2. Statement of the Baum-Connes conjecture

Let G be a second countable, locally compact group. We fix a left-invariant Haar measure dg on G. Denote by $C_c(G)$ the convolution algebra of complex-valued continuous compactly supported functions on G. The convolution of $f, f' \in C_c(G)$ is given by $f * f'(g) = \int_G f(h) f'(h^{-1}g) dh$ for any $g \in G$.

When G is discrete and dg is the counting measure, $C_c(G)$ is also denoted by $\mathbb{C}G$ and if e_g denotes the delta function at $g \in G$ (equal to 1 at g and 0 elsewhere), $(e_g)_{g \in G}$ is a basis of $\mathbb{C}G$ and the convolution product is given by $e_g e_{g'} = e_{gg'}$.

The completion of $C_c(G)$ for the norm $\|f\|_{L^1} = \int_G |f(g)| dg$ is a Banach algebra and is denoted by $L^1(G)$.

For any $f \in C_c(G)$ let $\lambda(f)$ be the operator $f' \mapsto f * f'$ on $L^2(G)$. The completion of $C_c(G)$ by the operator norm $\|f\|_{\mathrm{red}} = \|\lambda(f)\|_{\mathcal{L}_{\mathbb{C}}(L^2(G))}$ is called the reduced C^*-algebra of G and denoted by $C^*_{\mathrm{red}}(G)$. If G is discrete $(e_{g'})_{g' \in G}$ is an orthonormal basis of $L^2(G)$ and $\lambda(e_g) : e_{g'} \mapsto e_{gg'}$.

For any $f \in C_c(G)$, $\|f\|_{L^1} \geq \|f\|_{\mathrm{red}}$ and $L^1(G)$ is a dense subalgebra of $C^*_{\mathrm{red}}(G)$. We denote by $i : L^1(G) \to C^*_{\mathrm{red}}(G)$ the inclusion.

Assume now that M is a smooth compact manifold, and $\tilde{M}$ a Galois covering of M with group G (if $\tilde{M}$ is simply connected, $G = \pi_1(M)$). Let E_0 and E_1 be two smooth hermitian finite-dimensional vector bundles over M and u an order 0 elliptic pseudo-differential operator from $L^2(M, E_0)$ to $L^2(M, E_1)$. Since u is elliptic there is an order 0 pseudo-differential operator $v : L^2(M, E_1) \to L^2(M, E_0)$ such that $\mathrm{Id}_{L^2(M,E_0)} - vu$ and $\mathrm{Id}_{L^2(M,E_1)} - uv$ have order ≤ -1 and therefore are compact. Let $\mathcal{E}$ be the quotient of $\tilde{M} \times C^*_{\mathrm{red}}(G)$ by the diagonal action of G (G acting on $C^*_{\mathrm{red}}(G)$ by left translations) : $\mathcal{E}$ is a flat bundle of right $C^*_{\mathrm{red}}(G)$-modules over M, whose fibers are isomorphic to $C^*_{\mathrm{red}}(G)$. Then $L^2(M, E_0 \otimes \mathcal{E})$ and $L^2(M, E_1 \otimes \mathcal{E})$ are right Banach (in fact Hilbert) modules over $C^*_{\mathrm{red}}(G)$ and it is possible to lift u and v to $\tilde{u}$ and $\tilde{v}$ so that $(L^2(M, E_0 \otimes \mathcal{E}), L^2(M, E_1 \otimes \mathcal{E}), \tilde{u}, \tilde{v})$ is a Fredholm module over $C^*_{\mathrm{red}}(G)$, whose index lies in $K_0(C^*_{\mathrm{red}}(G))$ and the index does not depend on the choice of the liftings.

The operator u represents a "K-homology class" in $K_0(M)$, and using the classifying map $M \to BG$, it defines an element of $K_{0,c}(BG)$, the K-homology with compact support of the classifying space BG. For any discrete group G we can define a morphism of abelian groups $K_{*,c}(BG) \to K_*(C^*_{\mathrm{red}}(G))$ $(* = 0, 1)$. This morphism is the Baum-Connes assembly map when G is discrete and torsion free. When G is not discrete or has torsion, the index construction can be performed starting from a proper action of G (instead of the free and proper action of G on $\tilde{M}$ in the last paragraph), and therefore we have to introduce the space $\underline{E}G$ that classifies the proper actions of G. Using Kasparov equivariant KK-theory, the G-equivariant K-homology $K^G_*(\underline{E}G)$ with G-compact support $(* = 0, 1)$ may be defined, and there is an assembly map

$$\mu_{\mathrm{red}} : K^G_*(\underline{E}G) \to K_*(C^*_{\mathrm{red}}(G)).$$

In the same way we can define $\mu_{L^1} : K^G_*(\underline{E}G) \to K_*(L^1(G))$ and $\mu_{\mathrm{red}} = i_* \circ \mu_{L^1}$.

Baum-Connes conjecture [3, 4] : *If G is a second countable, locally compact group then the assembly map $\mu_{\mathrm{red}} : K^G_*(\underline{E}G) \to K_*(C^*_{\mathrm{red}}(G))$ is an isomorphism.*

Bost conjectured : If G is a second countable, locally compact group (and has reasonable geometric properties) then the assembly map $\mu_{L^1} : K^G_*(\underline{E}G) \to K_*(L^1(G))$ is an isomorphism.

In many cases $K^G_*(\underline{E}G)$ can be computed. For instance if G is a discrete torsion free subgroup of a reductive Lie group H and K is a maximal compact subgroup of H, then a possible $\underline{E}G$ is H/K and $K^G_*(\underline{E}G)$ is the K-homology with compact support of $G\backslash H/K$. This group may be computed thanks to Mayer Vietoris sequences. See part 2 for the case where G is a Lie group.

1.3. KK-theory

For any C^*-algebras A and B, Kasparov [28, 31] defined an abelian group $KK(A, B)$, covariant in B and contravariant in A. There is a product $KK(A, B) \otimes$

$KK(B,C) \to KK(A,C)$. Moreover $KK(\mathbb{C},A) = K_0(A)$ and therefore the product gives a morphism $KK(A,B) \to \mathrm{Hom}(K_0(A),K_0(B))$. The definition of $KK(A,B)$ is like definition 1.3.1 below, but with Hilbert modules instead of Banach modules.

For any Banach algebras A and B, we define [37, 49] an abelian group $KK^{\mathrm{ban}}(A,B)$, covariant in B and contravariant in A. There is no product, but a morphism $KK^{\mathrm{ban}}(A,B) \to \mathrm{Hom}(K_0(A),K_0(B))$. Assume that B is unital (otherwise the definition has to be slightly modified).

Definition 1.3.1 *$E^{\mathrm{ban}}(A,B)$ is the set of isomorphism classes of data (E,π,T), where E is a $\mathbb{Z}/2\mathbb{Z}$-graded right Banach module, $\pi : A \to \mathcal{L}_B(E)$ is a morphism of Banach algebras and takes values in even operators, and $T \in \mathcal{L}_B(E)$ is odd and satisfies $a(T^2 - \mathrm{Id}_E) \in K_B(E)$ and $aT - Ta \in K_B(E)$ for any $a \in A$.*

Then $KK^{\mathrm{ban}}(A,B)$ is the set of homotopy classes in $E^{\mathrm{ban}}(A,B)$, where the homotopy relation is defined using $E^{\mathrm{ban}}(A,B[0,1])$.

Remark : $E^{\mathrm{ban}}(\mathbb{C},B)$ is the set of isomorphism classes of Fredholm modules over B and $KK^{\mathrm{ban}}(\mathbb{C},B) = K_0(B)$.

If p is an idempotent in A, and $(E,\pi,T) \in E^{\mathrm{ban}}(A,B)$, the image of $[p] \in K_0(A)$ by the image of $[(E,\pi,T)] \in KK^{\mathrm{ban}}(A,B)$ in $\mathrm{Hom}(K_0(A),K_0(B))$ is defined to be the index of the Fredholm module over B equal to $(\mathrm{Im}\pi(p),\pi(p)T\pi(p))$. When p is an idempotent in $M_k(A)$, we use the image of p by $M_k(A) \to \mathcal{L}_B(E^k)$. This is enough to define the morphism $KK^{\mathrm{ban}}(A,B) \to \mathrm{Hom}(K_0(A),K_0(B))$, when A is unital.

The same definition with ungraded modules gives $KK_1^{\mathrm{ban}}(A,B)$, and, with the notation $KK = KK_0$, we have a morphism $KK_i^{\mathrm{ban}}(A,B) \to \mathrm{Hom}(K_j(A),K_{i+j}(B))$, where all the indices are modulo 2.

1.4. Status of injectivity and the element γ

The injectivity of the Baum-Connes map μ_{red} (and therefore of μ_{L^1}) is known for the following very large classes of groups :

a) groups acting continuously properly isometrically on a complete simply connected riemannian manifold with controlled non-positive sectional curvature, and in particular closed subgroups of reductive Lie groups ([29, 31]),

b) groups acting continuously properly isometrically on an affine building and in particular closed subgroups of reductive p-adic groups ([32]),

c) groups acting continuously properly isometrically on a discrete metric space with good properies at infinity (weakly geodesic, uniformly locally finite, and "bolic" [33, 34]), and in particular hyperbolic groups (*i.e.* word-hyperbolic in the sense of Gromov),

d) groups acting continuously amenably on a compact space ([22]).

In the cases a),b),c) above, the proof of injectivity provides an explicit idempotent endomorphism on $K_*(C^*_{\mathrm{red}}(G))$ whose image is the image of μ_{red} (and the

same for μ_{L^1}). In case d), J.-L. Tu has also constructed such an endomorphism, but in a less explicit way.

To state this we need to understand a baby case of Kasparov's equivariant KK-groups. Let G be a second countable, locally compact group. We denote by $E_G(\mathbb{C},\mathbb{C})$ the set of isomorphism classes of triples (H,π,T) where H is a $\mathbb{Z}/2$-graded Hilbert space, π a unitary representation of G on H (such that for any $x \in H$, $g \mapsto gx$ is continuous from G to H) and T an odd operator on H such $T^2 - \mathrm{Id}_H$ is compact and $\pi(g)T\pi(g^{-1}) - T$ is compact and depends norm continuously on $g \in G$. Then $KK_G(\mathbb{C},\mathbb{C})$ is the quotient of $E_G(\mathbb{C},\mathbb{C})$ by homotopy. Kasparov proved that $KK_G(\mathbb{C},\mathbb{C})$ has a ring structure (using direct sum for the addition and tensor products together with a quite difficult construction for the multiplication).

If π is a unitary representation of G on a Hilbert space H_0 and $H_1 = 0$ then $(H,\pi,0) \in E_G(\mathbb{C},\mathbb{C})$ if and only if H_0 has finite dimension. If moreover $H_0 = \mathbb{C}$ and π is the trivial representation of G, the class of $(H,\pi,0)$ is the unit of $KK_G(\mathbb{C},\mathbb{C})$ and is denoted by 1. If G is compact the classes of $(H,\pi,0)$ with $H_1 = 0$ (and $\dim H_0 < +\infty$) generate $KK_G(\mathbb{C},\mathbb{C})$ and $KK_G(\mathbb{C},\mathbb{C})$ is equal to the representation ring of G.

The important fact is that there is a "descent morphism"

$$j_{\mathrm{red}} : KK_G(\mathbb{C},\mathbb{C}) \to \mathrm{End}(K_*(C^*_{\mathrm{red}}(G))).$$

In fact it is a ring homomorphism and $j_{\mathrm{red}}(1) = \mathrm{Id}_{K_*(C^*_{\mathrm{red}}(G))}$. It is defined as the composite of two maps $KK_G(\mathbb{C},\mathbb{C}) \to KK(C^*_{\mathrm{red}}(G), C^*_{\mathrm{red}}(G)) \to \mathrm{End}(K_*(C^*_{\mathrm{red}}(G)))$. The construction of j_{red} is due to Kasparov. The construction of j_{L^1} to be explained below is an adaptation of it.

The following extremely important theorem also contains earlier works of Mishchenko and Solovjev.

Theorem 1.4.1 *(Kasparov, Kasparov-Skandalis [31, 32, 33, 34]) If G belongs to one of the classes a),b),c) above, the geometric conditions in a),b) or c) allow to construct an idempotent element $\gamma \in KK_G(\mathbb{C},\mathbb{C})$ such that μ_{red} is injective and its image is equal to the image of the idempotent $j_{\mathrm{red}}(\gamma) \in \mathrm{End}(K_*(C^*_{\mathrm{red}}(G)))$.*

1.5. Homotopies between γ and 1

We assume that G belongs to one of the classes a),b),c). Then the injectivity of μ_{red} is known and the surjectivity is equivalent to the equality $j_{\mathrm{red}}(\gamma) = \mathrm{Id} \in \mathrm{End}(K_*(C^*_{\mathrm{red}}(G)))$.

Theorem 1.5.1 *We have $\gamma = 1$ in $KK_G(\mathbb{C},\mathbb{C})$ if*

1. *G is a free group (Cuntz, [14]) or a closed subgroup of $SO(n,1)$ (Kasparov, [30]) or of $SU(n,1)$ (Julg-Kasparov, [25]) or of $SL_2(\mathbb{F})$ with $\mathbb{F}$ a local non-archimedian field (Julg-Valette, [24]),*
2. *G acts isometrically and properly on a Hilbert space (Higson-Kasparov [20, 27]).*

In fact the second case contains the first one.

If G has property (T) and is not compact, $\gamma \neq 1$ in $KK_G(\mathbb{C},\mathbb{C})$: it is impossible to deform 1 to γ in $E_G(\mathbb{C},\mathbb{C})$ because the trivial representation is isolated among unitary representations of G if G has property (T) and γ can be represented by (H,π,T) such that H has no invariant vector (and even H is tempered). All simple real or p-adic groups of rank ≥ 2, and $Sp(n,1)$ and $F_{4(-20)}$, and all their lattices, have property (T) (see [19]).

It is then natural to broaden the class of representations in order to break the isolation of the trivial one. In [26] Julg proposed to use uniformly bounded representations on Hilbert spaces (to solve the case of $Sp(n,1)$).

For any non compact group G the trivial representation is not isolated among isometric representations in Banach spaces (think of the left regular representation on $L^p(G)$, p going to infinity).

Definition 1.5.2 *Let $E_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C})$ be the set of isomorphism classes of triples (E,π,T) with E a $\mathbb{Z}/2$-graded Banach space endowed with an isometric representation of G (such that $g \mapsto gx$ is continuous from G to E for any $x \in E$), $T \in \mathcal{L}_{\mathbb{C}}(E)$ an odd operator such that $T^2 - \mathrm{Id}_E$ belongs to $\mathcal{K}_{\mathbb{C}}(E)$ and $\pi(g)T\pi(g^{-1}) - T$ belongs to $\mathcal{K}_{\mathbb{C}}(E)$ and depends norm continuously on $g \in G$.*

Then $KK_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C})$ is defined as the quotient of $E_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C})$ by homotopy.

Since any unitary representation of G on a Hilbert space H is an isometric representation on the Banach space H, there is a natural morphism of abelian groups $KK_G(\mathbb{C},\mathbb{C}) \to KK_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C})$.

To state our main theorem, we need to look at slightly smaller classes than a) and c) above. We call these new classes a') and c'). They are morally the same, and in particular they respectively contain all closed subgroups of reductive Lie groups, and all hyperbolic groups (for general hyperbolic groups see [42], and [37] for a slightly different approach).

Theorem 1.5.3 *[37, 49] For any group G in the classes a'), b), or c'), we have $\gamma = 1$ in $KK_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C})$.*

In fact the statement is slightly incorrect, we should allow representations with a slow growth, but this adds no real difficulty. The proof of this theorem is quite technical. Let me just indicate some ingredients involved. If G is in class a') then G acts continuously isometrically properly on a complete simply connected riemmannian manifold X with controlled non-positive sectional curvature, and X is contractible (through geodesics) and the de Rham cohomology of X (without support) is $\mathbb{C}$ in degree 0 and 0 in other degrees. It is possible to put norms on the spaces of differential forms (on which G acts) and to build a parametrix for the de Rham operator (in the spirit of the Poincaré lemma) in order to obtain a resolution of the trivial representation, and in our language an element of $E_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C})$ equal to 1 in $KK_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C})$. The norms we use are essentially Sobolev L^∞ norms. Then it is possible to conjugate the operators by an exponential of the distance to a fixed point in X and then to deform these norms to Hilbert norms (through L^p norms, $p \in [2,+\infty]$) and to reach γ.

If G belongs to class b) the de Rham complex is replaced by the simplicial homology complex (with L^1 norms) on the building. If G belongs to class c') a Rips complex plays the same role as the building in b).

It is not possible to apply directly this theorem to the Baum-Connes conjecture because there is no obvious descent map $KK_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C}) \to \mathrm{End}(K_*(C^*_{\mathrm{red}}(G)))$, and in the next subsection we shall see the difficulties encountered and the way one bypasses them in a few cases.

On the other hand, we may apply this theorem to Bost conjecture, because there is descent map $j_{L^1} : KK_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C}) \to KK^{\mathrm{ban}}(L^1(G),L^1(G))$.

We explain it when G is discrete. Let $(E,\pi,T) \in E_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C})$. We denote by $L^1(G,E)$ the completion of $E \otimes \mathbb{C}G$ for the norm $\|\sum_{g\in G} x(g)\otimes e_g\| = \sum_{g\in G}\|x(g)\|_E$. Then $L^1(G,E)$ is a right Banach $L^1(G)$-module by the formula $(x\otimes e_g)e_{g'} = x\otimes e_{gg'}$ and there is a Banach algebra morphism $\hat{\pi} : L^1(G) \to \mathcal{L}_{L^1(G)}(L^1(G,E))$ by the formula $\hat{\pi}(e_{g'})(x\otimes e_g) = \pi(g')(x)\otimes e_{g'g}$. Then $(L^1(G,E),\hat{\pi}, T\otimes 1) \in E^{\mathrm{ban}}(L^1(G),L^1(G))$ gives the desired class in $KK^{\mathrm{ban}}(L^1(G),L^1(G))$.

This and section 1.3 imply the Bost conjecture in many cases.

Theorem 1.5.4 *For any group G in the classes a'), b) or c'), $\mu_{L^1} : K_*^G(\underline{E}G) \to K_*(L^1(G))$ is an isomorphism.*

1.6. Unconditional completions

Let G be a second countable, locally compact group. Let $\mathcal{A}(G)$ be a Banach algebra containing $C_c(G)$ as a dense subalgebra. We write $\mathcal{A}(G)$ instead of $\mathcal{A}$ for notational convenience. We ask for a necessary and sufficient condition such that there is a "natural" descent map $j_{\mathcal{A}} : KK_G^{\mathrm{ban}}(\mathbb{C},\mathbb{C}) \to KK^{\mathrm{ban}}(\mathcal{A}(G),\mathcal{A}(G))$.

In order to simplify the argument below, we will assume G to be discrete.

Let E be a Banach space with an isometric representation of G. Then $E\otimes\mathbb{C}G$ has a right $\mathbb{C}G$-module structure given by $(x\otimes e_g)e_{g'} = x\otimes e_{gg'}$ and there is a morphism $\hat{\pi} : \mathbb{C}G \to \mathrm{End}_{\mathbb{C}G}(E\otimes\mathbb{C}G)$ given by the formula $\hat{\pi}(e_{g'})(x\otimes e_g) = \pi(g')(x)\otimes e_{g'g}$. We look for a completion $\mathcal{A}(G,E)$ of $E\otimes\mathbb{C}G$ by a Banach norm such that $\mathcal{A}(G,E)$ is a right Banach $\mathcal{A}(G)$-module and $\hat{\pi}$ extends to a morphism of Banach algebras $\hat{\pi} : \mathcal{A}(G) \to \mathcal{L}_{\mathcal{A}(G)}(\mathcal{A}(G,E))$.

In order to have enough $\mathcal{A}(G)$-rank one operators, it is quite natural to assume that the norm on $\mathcal{A}(G,E)$ satisfies : for any $x\in E$ and $\xi\in\mathcal{L}_{\mathbb{C}}(E,\mathbb{C})$, if we denote by $R_x : \mathbb{C}G \to E\otimes\mathbb{C}G$ the map $e_g \mapsto x\otimes e_g$ and by $S_\xi : E\otimes\mathbb{C}G \to \mathbb{C}G$ the map $y\otimes e_g \mapsto \xi(y)e_g$, we have $\|R_x(f)\|_{\mathcal{A}(G,E)} \le \|x\|_E\|f\|_{\mathcal{A}(G)}$ for any $f\in\mathbb{C}G$ and $\|S_\xi(\omega)\|_{\mathcal{A}(G)} \le \|\xi\|_{\mathcal{L}_{\mathbb{C}}(E,\mathbb{C})}\|\omega\|_{\mathcal{A}(G,E)}$ for any $\omega\in E\otimes\mathbb{C}G$. Now fix $x\in E$ and $\xi\in\mathcal{L}_{\mathbb{C}}(E,\mathbb{C})$ and denote by 1 the unit in G. For any $f = \sum_{g\in G} f(g)e_g \in \mathbb{C}G$, $S_\xi(\hat{\pi}(f)(R_x(e_1)))$ is $\sum_{g\in G}\xi(\pi(g)(x))f(g)e_g$ in $\mathbb{C}G$. For any function c on G, we define the Schur multiplication by c to be the pointwise product $\mathbb{C}G \to \mathbb{C}G$, $\sum_{g\in G} f(g)e_g \mapsto \sum_{g\in G} c(g)f(g)e_g$.

In this way we obtain the following necessary condition : for any $x\in E$ and $\xi\in\mathcal{L}_{\mathbb{C}}(E,\mathbb{C})$ the Schur multiplication by the matrix coefficient $g\mapsto\xi(\pi(g)(x))$ is bounded from $\mathcal{A}(G)$ to itself and its norm (in $\mathcal{L}_{\mathbb{C}}(\mathcal{A}(G))$) is less than $\|x\|_E\|\xi\|_{\mathcal{L}_{\mathbb{C}}(E,\mathbb{C})}$. But for any L^∞-function c on G we can find an isometric representation π of G on

a Banach space E and $x \in E$ and $\xi \in \mathcal{L}_{\mathbb{C}}(E, \mathbb{C})$ such that $\|x\|_E \|\xi\|_{\mathcal{L}_{\mathbb{C}}(E,\mathbb{C})} = \|c\|_{L^\infty}$ and $c(g) = \xi(\pi(g)x)$ for any $g \in G$ (take $E = L^1(G)$, $x = \delta_1$, $\xi = c$). Therefore a necessary condition is that $\mathcal{A}(G)$ is an unconditional completion in the following sense.

Definition 1.6.1 *A Banach algebra $\mathcal{A}(G)$ (with a given norm $\|.\|_{\mathcal{A}(G)}$) containing $C_c(G)$ as a dense subalgebra is called an unconditional completion if the norm $\|f\|_{\mathcal{A}(G)}$ of $f \in C_c(G)$ only depends on $g \mapsto |f(g)|$, $G \to \mathbb{R}_+$.*

Remark that $L^1(G)$ is an unconditional completion of $C_c(G)$ but $C^*_{\mathrm{red}}(G)$ is not.

In fact this condition is also sufficient to construct the descent map. For the sake of simplicity, we still assume that G is discrete. If $\mathcal{A}(G)$ is an unconditional completion of $\mathbb{C}G$, and (E, π, T) is in $E^{\mathrm{ban}}_G(\mathbb{C}, \mathbb{C})$, we define $\mathcal{A}(G, E)$ as the completion of $E \otimes \mathbb{C}G$ for the norm $\|\sum_{g\in G} x(g) \otimes e_g\| = \|\sum_{g\in G} \|x(g)\|_E \, e_g\|_{\mathcal{A}(G)}$ and $\mathcal{A}(G, E)$ is a right Banach module over $\mathcal{A}(G)$ and there is a morphism $\hat{\pi} : \mathcal{A}(G) \to \mathcal{L}_{\mathcal{A}(G)}(\mathcal{A}(G, E))$, and $(\mathcal{A}(G, E), \hat{\pi}, T \otimes 1) \in E^{\mathrm{ban}}(\mathcal{A}(G), \mathcal{A}(G))$.

In this way, for any unconditional completion $\mathcal{A}(G)$ of $\mathbb{C}G$, we have a descent map $j^{\mathrm{ban}}_{\mathcal{A}} : KK^{\mathrm{ban}}_G(\mathbb{C}, \mathbb{C}) \to KK^{\mathrm{ban}}(\mathcal{A}(G), \mathcal{A}(G)) \to \mathrm{End}(K_*(\mathcal{A}(G)))$. We can also define an assembly map $\mu_{\mathcal{A}} : K^G_*(\underline{E}G) \to K_*(\mathcal{A}(G))$. If $\mathcal{A}(G)$ is an involutive subalgebra of $C^*_{\mathrm{red}}(G)$, and $i : \mathcal{A}(G) \to C^*_{\mathrm{red}}(G)$ denotes the inclusion, $\mu_{\mathrm{red}} = i_* \circ \mu_{\mathcal{A}}$.

Theorem 1.6.2 *([37]) For any group G in the classes a'), b) or c'), and for any unconditional completion $\mathcal{A}(G)$ of $C_c(G)$, $\mu_{\mathcal{A}} : K^G_*(\underline{E}G) \to K_*(\mathcal{A}(G))$ is an isomorphism.*

Let A, B be Banach algebras and $i : A \to B$ an injective morphism of Banach algebras. We say that A is stable under holomorphic functional calculus in B if any element of A has the same spectrum in A and in B. If A is dense and stable under holomorphic functional calculus in B then $i_* : K_*(A) \to K_*(B)$ is an isomorphism (see the appendix of [6]).

Corollary 1.6.3 *For any group G in the classes a'), b) or c'), if $C_c(G)$ admits an unconditional completion $\mathcal{A}(G)$ which is an involutive subalgebra of $C^*_{\mathrm{red}}(G)$ and is stable under holomorphic functional calculus in $C^*_{\mathrm{red}}(G)$, then $\mu_{\mathrm{red}} : K^G_*(\underline{E}G) \to K_*(C^*_{\mathrm{red}}(G))$ is an isomorphism.*

This condition is fulfilled for

a) hyperbolic groups,

b) cocompact lattices in a product of a finite number of groups among Lie or p-adic groups of rank one, $SL_3(\mathbb{F})$ with $\mathbb{F}$ a local field (even $\mathbb{H}$) and $E_{6(-26)}$,

c) reductive Lie groups and reductive groups over non-archimedian local fields.

In case c), $\mathcal{A}(G)$ is a variant of the Schwartz algebra of the group ([37]). In this case the Baum-Connes conjecture was already known for linear connected reductive groups (Wassermann [55]) and for the p-adic GL_n (Baum, Higson, Plymen [5]). In case a),b) this result is based on a property first introduced by Haagerup for the

free group and called (RD) (for rapid decay) by Jolissaint ([23]). In case a),b) G has property (RD) : this is due to Haagerup for free groups ([16]), Jolissant for "geometric hyperbolic groups", de la Harpe for general hyperbolic groups ([18]), Ramagge, Robertson and Steger for SL_3 of a non-archimedian local field ([47]), the author for $SL_3(\mathbb{R})$ and $SL_3(\mathbb{C})$ ([39]), Chatterji for $SL_3(\mathbb{H})$ and $E_{6(-26)}$ ([10]), and the remark that it holds for products is due to Ramagge, Robertson and Steger ([47]) in a particular case, and independantly to Chatterji ([10]) and Talbi ([50]) in general. A discrete group G has property (RD) if there is a lenght function $\ell : G \to \mathbb{R}_+$ (*i.e.* a function satisfying $\ell(g^{-1}) = \ell(g)$ and $\ell(gh) \leq \ell(g) + \ell(h)$ for any $g, h \in G$) such that for $s \in \mathbb{R}_+$ big enough, the completion $H^s(G)$ of $\mathbb{C}G$ for the norm $\|\sum f(g)e_g\|_{H^s(G)} = \|\sum (1+\ell(g))^s f(g)e_g\|_{L^2(G)}$ is contained in $C^*_{\mathrm{red}}(G)$. Then, for s big enough, $H^s(G)$ is a Banach algebra and an involutive subalgebra of $C^*_{\mathrm{red}}(G)$ and is dense and stable under holomorphic functional calculus ([23, 39]); it is obvious that $H^s(G)$ is an unconditional completion of $\mathbb{C}G$.

As a consequence of this result the Baum-Connes conjecture has been proven for all almost connected groups by Chabert, Echterhoff and Nest ([9]).

1.7. Trying to push the method further

In order to prove new cases of the surjectivity of the Baum-Connes map (when the injectivity is proven and the γ element exists) we should look for a dense subalgebra $\mathcal{A}(G)$ of $C^*_{\mathrm{red}}(G)$ that is stable under holomorphic functional calculus and a homotopy between γ and 1 through (perhaps special kind of) elements of $E^{\mathrm{ban}}_G(\mathbb{C}, \mathbb{C})$ which all give a map $K_*(\mathcal{A}(G)) \to K_*(C^*_{\mathrm{red}}(G))$ by the descent construction. Thanks to the discussion in subsection 1. a necessary condition for this is that for any (E, π, T) in the homotopy between γ and 1, for any $x \in E$ and $\xi \in \mathcal{L}_{\mathbb{C}}(E, \mathbb{C})$, the Schur multiplication by the matrix coefficient $g \mapsto \xi(\pi(g)(x))$ is bounded from $\mathcal{A}(G)$ to $C^*_{\mathrm{red}}(G)$ and has norm $\leq \|x\|_E \|\xi\|_{\mathcal{L}_{\mathbb{C}}(E,\mathbb{C})}$. So we should first look for a homotopy between γ and 1 such that the fewest possible matrix coefficients appear. For groups acting properly on buildings, this homotopy can be shown to exist. The problem for general discrete groups properly acting on buildings is to find a subalgebra $\mathcal{A}(G)$ of $C^*_{\mathrm{red}}(G)$ that is stable under holomorphic functional calculus and satisfies the condition with respect to these matrix coefficients. The first step (the crucial one I think) should be to find a subalgebra $\mathcal{A}(G)$ of $C^*_{\mathrm{red}}(G)$ that is stable under holomorphic functional calculus and satisfies the following condition : there is a integer n, a distance d on the building and a point x_0 on the building such that the Schur product by the characteristic function of $\{g \in G, d(x_0, gx_0) \leq r\}$ from $\mathcal{A}(G)$ to $C^*_{\mathrm{red}}(G)$ has norm less than $(1+r)^n$, for any $r \in \mathbb{R}_+$.

1.8. The Baum-Connes conjecture with coefficients

Let G be a second countable, locally compact group and A a G-Banach algebra (*i.e.* a Banach algebra on which G acts continuously by isometric automorphisms $g : a \mapsto g(a)$). The space $C_c(G, A)$ of A-valued continuous compactly supported functions on G is endowed with the following convolution product : $f * f'(g) = \int_G f(h)h(f'(h^{-1}g))dh$ and the completion $L^1(G, A)$ of $C_c(G, A)$ for the norm $\|f\| =$

$\int_G \|f(g)\|_A dg$ is a Banach algebra. More generally for any unconditional completion $\mathcal{A}(G)$, we define $\mathcal{A}(G,A)$ to be the completion of $C_c(G,A)$ for the norm $\|f\|_{\mathcal{A}(G,A)} = \left\|g \mapsto \|f(g)\|_A\right\|_{\mathcal{A}(G)}$.

For any G-Banach algebras A and B, we define in [37] an abelian group $KK_G^{\mathrm{ban}}(A,B)$. This is a contravariant functor in A and a covariant functor in B. When $G=1$ this is equal to $KK^{\mathrm{ban}}(A,B)$. For any unconditional completion $\mathcal{A}(G)$ of $C_c(G)$, there is descent morphism $KK_G^{\mathrm{ban}}(A,B) \to KK^{\mathrm{ban}}(\mathcal{A}(G,A),\mathcal{A}(G,B))$.

These constructions are adaptations of the classical constructions for C^*-algebras : for any G-C^*-algebra A (*i.e.* G acts continuously by C^*-algebras automorphisms on A) we have a natural C^*-algebra $C^*_{\mathrm{red}}(G,A)$ containing $L^1(G,A)$ as a dense subalgebra. If B is another G-C^*-algebra, Kasparov defined an abelian group $KK_G(A,B)$. This is a contravariant functor in A and a covariant functor in B. When $G=1$ this is equal to $KK(A,B)$. There is an associative and distributive product $KK_G(A,B) \otimes KK_G(B,C) \to KK_G(A,C)$ and a descent morphism $KK_G(A,B) \to KK(C^*_{\mathrm{red}}(G,A), C^*_{\mathrm{red}}(G,B))$.

Let $K_*^G(\underline{E}G,A)$, $*=0,1$, be the inductive limit over G-invariant G-compact subsets Z of $\underline{E}G$ of $KK_{G,*}(C_0(Z),A)$. Then the assembly map

$$\mu_{\mathrm{red},A} : K_*^G(\underline{E}G,A) \to K_*(C^*_{\mathrm{red}}(G,A))$$

is defined in [4] and similar maps $\mu_{L^1,A}$, and more generally $\mu_{\mathcal{A},A}$ for any unconditional completion $\mathcal{A}(G)$, can be defined.

The Baum-Connes conjecture "with coefficients" claims that $\mu_{\mathrm{red},A}$ is an isomorphism and the Bost conjecture "with coefficients" claims that $\mu_{L^1,A}$ is an isomorphism. Theorems 1.4.1, 1.5.4, 1.6.2 are still true with arbitrary coefficients.

The surjectivity of the Baum-Connes conjecture with coefficients has been counter-exampled recently (Higson, Lafforgue, Ozawa, Skandalis, Yu) using a random group constructed by Gromov ([15]) but Bost conjecture with coefficients still stands. If the Baum-Connes conjecture with coefficients is true for a group, it is true also for all its closed subgroups; the Baum-Connes conjecture with coefficients is also stable under various kinds of extensions (Chabert [7], Chabert-Echterhoff [8], Oyono [44], and Tu [51]).

Kasparov's equivariant KK-theory was generalized to groupoids by Le Gall [31, 40, 41] and this generalized KK-theory was applied by Tu in [52, 53] to the bijectivity of the Baum-Connes map for amenable groupoids and the injectivity for (the holonomy groupoids of) hyperbolic foliations. It is possible to generalize also Banach KK-theory and unconditional completions. In this way we obtain the Baum-Connes conjecture for any hyperbolic group, with coefficients in any commutative C^*-algebra, and also for foliations with compact basis, admitting a (strictly) negatively curved longitudinal riemannian metric, and such that the holonomy groupoid is Hausdorff and has simply connected fibers (not yet published).

2. Discrete series representations of connected semi-simple Lie groups

In this part we examine how the Baum-Connes conjecture for a connected semi-simple Lie group with finite center can be used to establish the construction of the discrete series by Dirac induction ([17, 45, 1]). That this is morally true is known from the beginning of the conjecture (see for instance [12]). In the proof we shall introduce 3 ingredients : these are classical facts stated here without proof. Parts of the argument apply to more general groups (not connected, not semi-simple).

This work owes its existence to Paul Baum. He asked me to study the problem and we discussed a lot.

2.1. Dirac operators

Let G be a Lie group, with a finite number of connected components, and K a maximal compact subgroup. We assume that there exists a G-invariant orientation on G/K. For the sake of simplicity, we assume that G/K admits a G-invariant spin structure (it is true anyway for a two fold covering of G). More precisely let $\mathfrak{p}$ be a complementary subspace for the Lie algebra $\mathfrak{k}$ of K in the Lie algebra $\mathfrak{g}$ of G. We choose $\mathfrak{p}$ such that it is invariant for the adjoint action of K and we endow it with a K-invariant euclidian metric. The above assumption means that the homomorphism $K \to \mathrm{SO}(\mathfrak{p})$ lifts to $\mathrm{Spin}(\mathfrak{p})$. We denote by S the associated spin representation of K. If $\dim(G/K)$ is even, S is $\mathbb{Z}/2\mathbb{Z}$-graded. We write $i = \dim(G/K) \ [2]$.

We denote by $R(K)$ the (complex) representation ring of K and for any finite dimensional representation V of K we denote by $[V]$ its class in $R(K)$.

Let V be a finite dimensional representation of K. Let E_V be the right Banach (in fact Hilbert) module over $C^*_{\mathrm{red}}(G)$ ($\mathbb{Z}/2\mathbb{Z}$-graded if $i = 0$ [2]) whose elements are the K-invariant elements in $V^* \otimes S^* \otimes C^*_{\mathrm{red}}(G)$, where K acts by left translations on $C^*_{\mathrm{red}}(G)$. Let D_V be the unbounded $C^*_{\mathrm{red}}(G)$-linear operator on E_V equal to $\sum 1 \otimes c(p_i) \otimes p_i$, where the sum is over i, (p_i) is an orthonormal basis of $\mathfrak{p}$, p_i denotes also the associated right invariant vector field on G, and $c(p_i)$ is the Clifford multiplication by p_i. Let $T_V = \frac{D_V}{\sqrt{1+D_V^2}}$. Then we define $[d_V] \in K_i(C^*_{\mathrm{red}}(G))$ to be the class of the Fredholm module (E_V, T_V) over $C^*_{\mathrm{red}}(G)$.

In other words, E_V is the completion of the space of smooth compactly supported sections of the bundle on $K\backslash G$ associated to the representation $V^* \otimes S^*$ of K, for the norm $\|w\| = \sup_{f \in L^2(G), \|f\|_{L^2(G)}=1} \|w * f\|_{L^2((V^* \otimes S^*) \times_K G)}$, and D_V is the Dirac operator, twisted by V^*.

Connes-Kasparov conjecture. *The group morphism $\mu_{\mathrm{red}} : R(K) \to K_i(C^*_{\mathrm{red}}(G))$ defined by $[V] \mapsto [d_V]$ is an isomomorphism, and $K_{i+1}(C^*_{\mathrm{red}}(G)) = 0$.*

This is a special case of the Baum-Connes conjecture because we may take $\underline{E}G = G/K$ and thus $K_i^G(\underline{E}G) = R(K)$ and $K_{i+1}^G(\underline{E}G) = 0$. It was checked for G connected reductive linear in [55] and the Baum-Connes conjecture was proved for any reductive group in [37] (see c) of the corollary 1.6.3 above).

The following lemma has been suggested to me by Francois Pierrot. Assume that i is even. Let moreover H be a unitary tempered admissible representation of G. This implies that we have a C^*-homomorphism $C^*_{\mathrm{red}}(G) \to \mathcal{K}(H)$. For any element $x \in K_0(C^*_{\mathrm{red}}(G))$ we denote by $\langle H, x \rangle \in \mathbb{Z}$ the image of x by $K_0(C^*_{\mathrm{red}}(G)) \to K_0(\mathcal{K}(H)) = \mathbb{Z}$. If x is the class of an idempotent $p \in C^*_{\mathrm{red}}(G)$, the image of p in

$\mathcal{K}(H)$ is a finite rank projector, whose rank is $\langle H, x\rangle$.

Lemma 2.1.1 *We have* $\langle H, [d_V]\rangle = \dim(V^* \otimes S^* \otimes H)^K$.

2.2. Dual-Dirac operators

From now on we assume that G is a connected semi-simple Lie group with finite center and we still assume that G/K has a G-invariant spin structure. Kasparov has constructed an element $\eta \in \mathrm{Hom}(K_i(C^*_{\mathrm{red}}(G)), R(K))$ (coming from an element of $KK_i(C^*_{\mathrm{red}}(G), C^*_{\mathrm{red}}(K))$, itself coming from an element of $KK_{G,i}(\mathbb{C}, C_0(G/K))$). Kasparov has shown that $\eta \circ \mu_{\mathrm{red}} = \mathrm{Id}_{R(K)}$ [29, 31].

Here is the detail of the construction. The G-invariant riemannian structure on G/K given by the chosen K-invariant euclidian metric on $\mathfrak{p}$ has non-positive curvature. Let ρ be the distance to the origin and $\xi = d(\sqrt{1+\rho^2})$. Let V be a finite dimensional complex representation of K, endowed with an invariant hermitian metric. Let H_V be the space of L^2 sections of the hermitian G-equivariant fibre bundle on G/K associated to the representation of K on $S \otimes V$ and let $c_{\xi,V}$ be the Clifford multiplication by ξ. In other words H_V is the subspace of K-invariant vectors in $L^2(G) \otimes S \otimes V$, where K acts by right translations $L^2(G)$, and $c_{\xi,V}$ is the restriction to this subspace of the tensor product of the Clifford multiplication by ξ on $L^2(G) \otimes S$ with Id_V. Left translation by G on G/K or on $L^2(G)$ gives rise to a (C^*-)morphism $\pi_V : C^*_{\mathrm{red}}(G) \to \mathcal{L}_{\mathbb{C}}(H_V)$ and $(H_V, \pi_V, c_{\xi,V})$ defines $\eta_V \in KK_i^{\mathrm{ban}}(C^*_{\mathrm{red}}(G), \mathbb{C})$ (in fact in $KK_i(C^*_{\mathrm{red}}(G), \mathbb{C})$). We denote by $[\eta_V] \in \mathrm{Hom}(K_i(C^*_{\mathrm{red}}(G)), \mathbb{Z})$ the associated map, and $\eta = \sum_V [\eta_V][V] \in \mathrm{Hom}(K_i(C^*_{\mathrm{red}}(G)), R(K))$, where the sum is over the irreducible representations of K.

Since the Connes-Kasparov conjecture is true, $\mu_{\mathrm{red}} : R(K) \to K_i(C^*_{\mathrm{red}}(G))$ and $\eta : K_i(C^*_{\mathrm{red}}(G)) \to R(K)$ are inverse of each other and $K_{i+1}(C^*_{\mathrm{red}}(G)) = 0$.

Let H be a discrete series representation of G, *i.e.* an irreducible unitary representation with a positive mass in the Plancherel measure. We recall that this is equivalent to the fact that some (whence all) matrix coefficient $c_x(g) = \langle x, \pi(g)x\rangle$, $x \in H$, $\|x\| = 1$, is square-integrable. Then $\|c_x\|^2_{L^2(G)}$ is independant of x, and its inverse is the formal degree d_H of H, which is also the mass of H in the Plancherel measure. We introduce a first ingredient.

Ingredient 1. All discrete series representations of G are isolated in the tempered dual.

In other words, all matrix coefficients belong to $C^*_{\mathrm{red}}(G)$. In fact a standard asymptotic expansion argument shows that for any K-finite vector $x \in H$, c_x belongs to the Schwartz algebra ([17], II, corollary 1 page 77).

Therefore there exists an idempotent $p \in C^*_{\mathrm{red}}(G)$ such that the image in $L^2(G)$ of the image of p by the left regular representation is H^* as a representation of G on the right. In fact we can take $p = d_H \overline{c_x}$ for any $x \in H$, $\|x\| = 1$, where $\overline{c_x}(g) = \overline{c_x(g)}$. The class of p in $K_0(C^*_{\mathrm{red}}(G))$ only depends on H and we denote it by $[H]$. It is easy to see that $i : \oplus_H \mathbb{Z} \to K_0(C^*_{\mathrm{red}}(G))$, $(n_H)_H \mapsto \sum_H n_H[H]$, where the sums are over the discrete series representations of G, is an injection. Indeed, if H and H' are discrete series representations of G, $\langle H', [H]\rangle = 1$ if $H = H'$ and 0 otherwise.

As a corollary we see that if $i = 1$ [2], G has no discrete series representations. From now on we assume $i = 0$ [2].

The first part of the following lemma was suggested to me by Georges Skandalis. Let H be a discrete series representation of G. We write $\eta([H]) = \sum_V n_V[V]$ in $R(K)$ where the sum is finite and over the irreducible representations of K (in the notation above, $n_V = [\eta_V]([H])$).

Lemma 2.2.1 *If V is an irreducible representation of K, $n_V = \dim(H^* \otimes S \otimes V)^K$ and therefore $n_V = \langle H, [d_V] \rangle$.*

We have $1 = \langle H, [H] \rangle = \langle H, \mu_{\text{red}} \circ \eta([H]) \rangle = \sum_V n_V \langle H, [d_V] \rangle = \sum_V n_V^2$. Therefore one of the n_V is ± 1 and the others are 0.

Alternatively we can consider the morphisms

$$\oplus_V \mathbb{Z}[V] = R(K) \overset{\mu_{\text{red}}}{\to} K_0(C^*_{\text{red}}(G)) \overset{\pi}{\to} \prod_H \mathbb{Z} \text{ where } \pi(x) = (\langle H, x \rangle)_H$$

$$\text{and} \qquad \oplus_H \mathbb{Z} \overset{i}{\to} K_0(C^*_{\text{red}}(G)) \overset{\eta}{\to} R(K) = \oplus_V \mathbb{Z}[V]$$

where the sums are over the irreducible representations V of K and the discrete series representations H of G. Their product $\pi \circ \mu_{\text{red}} \circ \eta \circ i = \pi \circ i$ is equal to the inclusion of $\oplus_H \mathbb{Z}$ in $\prod_H \mathbb{Z}$ and their matrices in the base $([V])_V$ and the canonical base of $\oplus_H \mathbb{Z}$ are transpose of each other. Therefore each column of the matrix of $\eta \circ i$ contains exactly one non-zero coefficient, which is equal to ± 1. A posteriori, π takes its values in $\oplus_H \mathbb{Z}$.

Corollary 2.2.2 *The discrete series representations of G are in bijection with a subset of the set of isomorphism classes of irreducible representations of K. The irreducible representation V of K associated to a discrete series representation H is such that $V = \pm(H \otimes S^*)$ as a formal combination of irreducible representations of K, and H occurs in the kernel of the twisted Dirac operator D_V.*

Corollary 2.2.3 *If $\operatorname{rank} G \neq \operatorname{rank} K$, G has no discrete series.*

In this case S^* is 0 in $R(K)$ (Barbasch and Moscovici [2] (1.2.5) page 156) : this was indicated to me by Henri Moscovici.

2.3. A trace formula

From now on we assume that $\operatorname{rank} G = \operatorname{rank} K$. Let T a maximal torus in K (therefore also in G). Choose a Weyl chamber for the root system of $\mathfrak{g}$ and choose the Weyl chamber of the root system of $\mathfrak{k}$ containing it. Let V be an irreducible representation of K, μ its highest wheight, and $\lambda = \mu + \rho_K$ where ρ_K is the half sum of the positive roots of $\mathfrak{k}$.

We recall that the unbounded trace $\operatorname{Tr} : C^*_{\text{red}}(G) \to \mathbb{R}, f \mapsto f(1)$ gives rise to a group morphism $K_0(C^*_{\text{red}}(G)) \to \mathbb{R}$. When H is a discrete series representation of G, $\operatorname{Tr}([H])$ is the value at 1 of $p = d_H \overline{c_x}$ for some $x \in H$, $\|x\| = 1$, and therefore it is the formal degree d_H of H and is > 0.

Ingredient 2. $\mathrm{Tr}([d_V]) = \prod_{\alpha\in\Psi} \frac{(\lambda,\alpha)}{(\rho,\alpha)}$, where Ψ is the set of simple roots of the chosen positive root system in $\mathfrak{g}$, and ρ is the half sum of the positive roots of this system.

In this formula is used a right normalization of the Haar measure (if G is linear it is the one for which the maximal compact subgroup of the complexification of G has measure 1). This formula is proven in [11] by a heat equation method, and in [1] by Atiyah's L^2-index theorem.

Corollary 2.3.1 *If λ is singular for $\mathfrak{g}$, $[V]$ does not correspond to a discrete series representation of G.*

Ingredient 3. For any $x \in K_0(C^*_{\mathrm{red}}(G))$ such that $\mathrm{Tr}(x) \neq 0$, there is a discrete series representation H such that $\langle H, x\rangle \neq 0$.

By the Plancherel formula, if $\hat{G}$ is the tempered spectrum of G, $\mathrm{Tr}(x) = \int_{\hat{G}} \langle H, x\rangle \, dH$. We have to prove that, for almost all H outside the discrete series, $\langle H, x\rangle = 0$. There are several possible arguments :

- almost all H outside the discrete series are induced from a parabolic subgroup and belong to a family of representations indexed by some $\mathbb{R}^p$, but $\langle H', x\rangle$ is constant when H' varies in this family and goes to 0 when H' goes to infinity,
- write $x = [d_V]$ for some V, then the H outside the discrete series with $\langle H, x\rangle \neq 0$ have measure 0 by [1] p15 (3.19), p50 (9.8) and p51 (9.12) or by [11] p318-320.

Corollary 2.3.2 *If λ is not singular for $\mathfrak{g}$, $[V]$ does correspond to a discrete series representation, whose formal degree is* $\left|\prod_{\alpha\in\Psi} \frac{(\lambda,\alpha)}{(\rho,\alpha)}\right|$.

We have recovered some results proved in [17], [45] and [1].

References

[1] M. Atiyah and W. Schmid, *A geometric construction of the discrete series for semisimple Lie groups*, Invent. Math. 42 (1977), 1–62.

[2] D. Barbasch and H. Moscovici, *L^2-index and the Selberg trace formula*, J. Funct. Anal. 53 (1983), no. 2, 151–201.

[3] P. Baum and A. Connes, *Geometric K-theory for Lie groups and foliations*, Preprint (1982), Enseign. Math. (2) **46** (2000), no. 1-2, 3–42.

[4] P. Baum and A. Connes and N. Higson, *Classifying space for proper actions and K-theory of group C^*-algebras*, C^*-algebras: 1943–1993 (San Antonio, TX, 1993), Contemp. Math., **167**, Amer. Math. Soc. (1994), 240–291.

[5] P. Baum, N. Higson and R. Plymen, *A proof of the Baum-Connes conjecture for p-adic $GL(n)$*, C. R. Acad. Sci. Paris Sér. I, **325,** (1997), 171–176.

[6] J.-B. Bost, *Principe d'Oka, K-théorie et systèmes dynamiques non commutatifs*, Invent. Math., **101** (1990), 261–333.

[7] J. Chabert, *Baum-Connes conjecture for some semi-direct products*, J. Reine Angew. Math., **521**, (2000) 161–184.

[8] J. Chabert and S. Echterhoff, *Permanence properties of the Baum-Connes conjecture*, Doc. Math. 6 (2001), 127–183.
[9] J. Chabert, S. Echterhoff and R. Nest, *The Connes-Kasparov conjecture for almost connected groups*, Preprint, University of Münster (2001).
[10] I. Chatterji, *Property (RD) for cocompact lattices in a finite product of rank one Lie groups with some rank two Lie groups*, to appear in Geometriae Dedicata.
[11] A. Connes and H. Moscovici, *The L^2-index theorem for homogeneous spaces of Lie groups*, Ann. of Math. (2) 115 (1982), no. 2, 291–330.
[12] A. Connes and H. Moscovici, *L^2-index theory on homogeneous spaces and discrete series representations*, Operator algebras and applications, Part I (Kingston, Ont., 1980), pp. 419–433, Proc. Sympos. Pure Math., 38, Amer. Math. Soc., Providence, R.I., 1982.
[13] A. Connes, *Non commutative geometry,* Academic Press, (1994).
[14] J. Cuntz, *K-theoretic amenability for discrete groups*, J. Reine Angew. Math. **344** (1983), 180–195.
[15] M. Gromov, *Spaces and questions,* Geom. Funct. Anal. 2000, Special Volume, Part I, 118–161.
[16] U. Haagerup, *An example of a nonnuclear C^*-algebra which has the metric approximation property,* Inv. Math., **50,** (1979), 279–293.
[17] Harish-Chandra, *Discrete series for semisimple Lie groups, I and II*, Acta Math. 113 (1965) 241–318 and 116 (1966), 1–111.
[18] P. de la Harpe, *Groupes hyperboliques, algèbres d'opérateurs et un théorème de Jolissaint,* C. R. Acad. Sci. Paris Sér. I, **307** (1988), 771–774.
[19] P. de la Harpe and A. Valette, *La propriété (T) de Kazdhan pour les groupes localement compacts,* Astérisque 175, (1989).
[20] N. Higson and G. Kasparov, *E-theory and KK-theory for groups which act properly and isometrically on Hilbert space.*, Invent. Math. 144, **1**, (2001), 23–74.
[21] N. Higson, *The Baum-Connes conjecture*, Proc. of the Int. Cong. of Math., Vol. II (Berlin, 1998), Doc. Math., (1998), 637–646.
[22] N. Higson, *Bivariant K-theory and the Novikov conjecture*, Geom. Funct. Anal. 10 (2000), no. 3, 563–581.
[23] P. Jolissaint, *Rapidly decreasing functions in reduced C^*-algebra of groups,* Trans. Amer. Math. Soc, **317,** (1990), 167–196.
[24] P. Julg and A. Valette, *K-theoretic amenability for $SL_2(Q_p)$, and the action on the associated tree,* J. Funct. Anal., **58**, (1984), 194–215.
[25] P. Julg and G. Kasparov, *Operator K-theory for the group $SU(n,1)$*, J. Reine Angew. Math., **463**, (1995), 99–152.
[26] P. Julg, *Remarks on the Baum-Connes conjecture and Kazhdan's property T*, Operator algebras and their applications, Waterloo (1994/1995), Fields Inst. Commun., Amer. Math. Soc. **13**, (1997), 145–153.
[27] P. Julg, *Travaux de N. Higson et G. Kasparov sur la conjecture de Baum-Connes*, Séminaire Bourbaki. Vol. 1997/98, Astérisque, **252**, (1998), No. 841, 4, 151–183.
[28] G. G. Kasparov, *The operator K-functor and extensions of C^*-algebras,* Math.

USSR Izv. **16** (1980), 513–572.

[29] G. G. Kasparov, *K-theory, group C^*-algebras, and higher signatures (conspectus)*, Novikov conjectures, index theorems and rigidity, Vol. 1 (Oberwolfach, 1993), 101–146, London Math. Soc. LNS 226, (1981).

[30] G. G. Kasparov *Lorentz groups: K-theory of unitary representations and crossed products*, Dokl. Akad. Nauk SSSR, **275**, (1984), 541–545.

[31] G. G. Kasparov, *Equivariant KK-theory and the Novikov conjecture*, Invent. Math. **91** (1988), 147–201.

[32] G. G. Kasparov and G. Skandalis, *Groups acting on buildings, operator K-theory, and Novikov's conjecture, K-*Theory, **4**, (1991), 303–337.

[33] G. Kasparov and G. Skandalis, *Groupes boliques et conjecture de Novikov,* Comptes Rendus Acad. Sc., **319** (1994), 815-820.

[34] G. Kasparov and G. Skandalis, *Groups acting properly on bolic spaces and the Novikov conjecture*, to appear in Annals of Math.

[35] V. Lafforgue, *Une démonstration de la conjecture de Baum-Connes pour les groupes réductifs sur un corps p-adique et pour certains groupes discrets possédant la propriété (T)*, C. R. Acad. Sci. Paris Sér. I, **327**, (1998), 439–444.

[36] V. Lafforgue, *Compléments à la démonstration de la conjecture de Baum-Connes pour certains groupes possédant la propriété (T)*, C. R. Acad. Sci. Paris Sér. I, **328**, (1999), 203–208.

[37] V. Lafforgue, *KK-théorie bivariante pour les algèbres de Banach et conjecture de Baum-Connes*, Invent. Math. 149 (2002) 1, 1-95.

[38] V. Lafforgue, *Banach KK-theory and the Baum-Connes conjecture*, European Congress of Mathematics (Barcelona 2000), Birkäuser, Volume 2.

[39] V. Lafforgue, *A proof of property (RD) for cocompact lattices of* $\mathrm{SL}(3,\mathbb{R})$ *and* $\mathrm{SL}(3,\mathbb{C})$, J. Lie Theory 10 (2000), no. 2, 255–267.

[40] P.-Y. Le Gall, *Théorie de Kasparov équivariante et groupoïdes*, C. R. Acad. Sci. Paris Sér. I, **324**, (1997), 695–698.

[41] P.-Y. Le Gall, *Théorie de Kasparov équivariante et groupoïdes. I, K-*Theory, **16**, (1999), 361–390.

[42] I. Mineyev and G. Yu, *The Baum-Connes conjecture for hyperbolic groups*, Invent. Math. 149 (2002) 1, 97-122

[43] A. S. Mishchenko, *Infinite-dimensional representations of discrete groups, and higher signatures*, Math. USSR Izv. **8**, no 1, (1974), 85–111.

[44] H. Oyono-Oyono, *Baum-Connes conjecture and group actions on trees*, K-Theory **24** (2001), no. 2, 115–134.

[45] R. Parthasarathy, *Dirac operator and the discrete series*, Ann. of Math. (2) 96 (1972), 1–30.

[46] M. V. Pimsner, *KK-groups of crossed products by groups acting on trees*, Invent. Math., **86**, (1986), 603–634.

[47] J. Ramagge, G. Robertson, T. Steger, *A Haagerup inequality for $\widetilde{A}_1 \times \widetilde{A}_1$ and $\widetilde{A}_2$ buildings*, Geom. Funct. Anal., **8** (1998), 702–731.

[48] G. Skandalis *Kasporov's bivariant K-theory and applications,* Expositiones Math., **9** (1991), 193–250.

[49] G. Skandalis *Progrès récents sur la conjecture de Baum-Connes, contribution*

de Vincent Lafforgue, Séminaire Bourbaki, No. 869 (novembre 1999).

[50] M. Talbi, *Inégalité de Haagerup et géométrie des groupes*, Thèse, Université de Lyon I (2001).

[51] J.-L. Tu, *The Baum-Connes conjecture and discrete group actions on trees*, K-Theory, **17**, (1999), 303–318.

[52] J.-L. Tu, *La conjecture de Baum-Connes pour les feuilletages moyennables*, K-Theory, **17**, (1999), 215–264.

[53] J.-L. Tu, *La conjecture de Novikov pour les feuilletages hyperboliques*, K-Theory, **16**, (1999), 129–184.

[54] A. Valette *An introduction to the Baum-Connes conjecture*, ETHZ Lectures in Mathematics, Birkhuser (2002).

[55] A. Wassermann, *Une démonstration de la conjecture de Connes-Kasparov pour les groupes de Lie linéaires connexes réductifs*, C. R. Acad. Sci. Paris Sér. I, **304** (1987), 559–562.

ICM 2002 · Vol. II · 813–822

On Some Inequalities for Gaussian Measures

R. Latała*

Abstract

We review several inequalities concerning Gaussian measures - isoperimetric inequality, Ehrhard's inequality, Bobkov's inequality, S-inequality and correlation conjecture.

2000 Mathematics Subject Classification: 60E15, 60G15, 28C20, 26D15.
Keywords and Phrases: Gaussian measure, Isoperimetry, Ehrhard's inequality, Convex bodies, Correlation.

1. Introduction

Gaussian random variables and processes always played a central role in the probability theory and statistics. The modern theory of Gaussian measures combines methods from probability theory, analysis, geometry and topology and is closely connected with diverse applications in functional analysis, statistical physics, quantum field theory, financial mathematics and other areas. Some examples of applications of Gaussian measures can be found in monographs [4, 18, 20] and [23].

In this note we present several inequalities of geometric nature for Gaussian measures. All of them have elementary formulations, but nevertheless yield many important and nontrivial consequences. We begin in section 2 with the already classical Gaussian isoperimetric inequality that inspired in the 70's and 80's the vigorous development of concentration inequalities and their applications in the geometry and local theory of Banach spaces (cf. [19, 24, 32]). In the sequel we review several more recent results and finish in section 6 with the discussion of the Gaussian correlation conjecture that remains unsolved more than 30 years.

A probability measure μ on a real separable Banach space F is called *Gaussian* if for every functional $x^* \in F^*$ the induced measure $\mu \circ (x^*)^{-1}$ is a one-dimensional Gaussian measure $\mathcal{N}(a, \sigma^2)$ for some $a = a(x^*) \in \mathbb{R}$ and $\sigma = \sigma(x^*) \geq 0$. Throughout this note we only consider centered Gaussian measures that is the measures such

*Institute of Mathematics, Warsaw University, Banacha 2, 02-097 Warszawa, Poland. E-mail: rlatala@mimuw.edu.pl

that $a(x^*) = 0$ for all $x^* \in F^*$. A random vector with values in F is said to be Gaussian if its distribution is Gaussian. Every centered Gaussian measure on $\mathbb{R}^n$ is a linear image of the canonical Gaussian measure γ_n, that is the measure on $\mathbb{R}^n$ with the density $d\gamma_n(x) = (2\pi)^{-n/2}\exp(-|x|^2/2)dx$, where $|x| = \sqrt{\sum_{i=1}^n x_i^2}$. Infinite dimensional Gaussian measures can be effectively approximated by finite dimensional ones using the following series representation (cf. [18, Proposition 4.2]): If μ is a centered Gaussian measure on F and $g_1, g_2, \ldots$ are independent $\mathcal{N}(0,1)$ random variables then there exist vectors $x_1, x_2, \ldots$ in F such that the series $X = \sum_{i=1}^\infty x_i g_i$ is convergent almost surely and in every L^p, $0 < p < \infty$, and is distributed as μ.

We will denote by Φ the distribution function of the standard normal $\mathcal{N}(0,1)$ r.v., that is

$$\Phi(x) = \gamma_1(-\infty, x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x e^{-y^2/2}dy, \ -\infty \le x \le \infty.$$

For two sets A, B in a Banach space F and $t \in \mathbb{R}$ we will write $tA = \{tx : x \in A\}$ and $A + B = \{x + y : x \in A, y \in B\}$. A set A in F is said to be *symmetric* if $-A = A$.

Many results presented in this note can be generalized to the more general case of Radon Gaussian measures on locally convex spaces. For precise definitions see [4] or [7].

2. Gaussian isoperimetry

For a Borel set A in $\mathbb{R}^n$ and $t > 0$ let $A_t = A + tB_2^n = \{x \in \mathbb{R}^n : |x - a| < t \text{ for some } a \in A\}$ be the open *t-enlargement* of A, where B_2^n denotes the open unit Euclidean ball in $\mathbb{R}^n$. The classical isoperimetric inequality for the Lebesgue measure states that if $\mathrm{vol}_n(A) = \mathrm{vol}_n(rB_2^n)$ then $\mathrm{vol}_n(A_t) \ge \mathrm{vol}_n((r+t)B_2^n)$ for $t > 0$. In the early 70's C. Borell [6] and V.N. Sudakov and B.S. Tsirel'son [29] proved independently the isoperimetric property of Gaussian measures.

Theorem 2.1 *Let A be a Borel set in $\mathbb{R}^n$ and let H be an affine halfspace such that $\gamma_n(A) = \gamma_n(H) = \Phi(a)$ for some $a \in \mathbb{R}$. Then*

$$\gamma_n(A_t) \ge \gamma_n(H_t) = \Phi(a + t) \text{ for all } t \ge 0. \tag{2.1}$$

Theorem 2.1 has an equivalent differential analog. To state it let us define for a measure μ on $\mathbb{R}^n$ and any Borel set A the *boundary μ-measure* of A by the formula

$$\mu^+(A) = \liminf_{t \to 0+} \frac{\mu(A_t) - \mu(A)}{t}.$$

Moreover let $\varphi(x) = \Phi'(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ and let

$$I(t) = \varphi \circ \Phi^{-1}(t), \ t \in [0,1]$$

be the *Gaussian isoperimetric function.*

The equivalent form of Theorem 2.1 is that for all Borel sets A in $\mathbb{R}^n$

$$\gamma_n^+(A) \geq I(\gamma_n(A)). \tag{2.2}$$

The equality in (2.2) holds for any affine halfspace.

For a probability measure μ on $\mathbb{R}^n$ we may define the *isoperimetric function* of μ by

$$\mathrm{Is}(\mu)(p) = \inf\{\mu^+(A) : \mu(A) = p\},\ 0 \leq p \leq 1.$$

Only few cases are known when one can determine exactly $\mathrm{Is}(\mu)$. For Gaussian measures (2.2) states that $\mathrm{Is}(\gamma_n) = I$.

Let us finish section 2 by an example of application of (2.1) (see [20, Lemma 3.1]).

Corollary 2.2 *Let X be a centered Gaussian random vector in a separable Banach space $(F, \|\cdot\|)$. Then for any $t > 0$*

$$\mathbf{P}(|\|X\| - \mathrm{Med}(\|X\|)| \geq t) \leq 2(1 - \Phi(\frac{t}{\sigma})) \leq e^{-t^2/2\sigma^2},$$

where

$$\sigma = \sup\{\sqrt{\mathbf{E}(x^*(X))^2} : x^* \in F^*, \|x^*\| \leq 1\}.$$

3. Ehrhard's inequality

It is well known that the classical isoperimetric inequality for the Lebesgue measure in $\mathbb{R}^n$ follows by the Brunn-Minkowski inequality (cf. [25]), which states that for any Borel sets A and B in $\mathbb{R}^n$

$$\mathrm{vol}_n(\lambda A + (1-\lambda)B)) \geq (\mathrm{vol}_n(A))^\lambda (\mathrm{vol}_n(B))^{1-\lambda} \text{ for } \lambda \in [0,1].$$

Gaussian measures satisfy the similar log-concavity property, that is the inequality

$$\ln(\mu(\lambda A + (1-\lambda)B)) \geq \lambda \ln(\mu(A)) + (1-\lambda)\ln(\mu(B)),\ \lambda \in [0,1] \tag{3.1}$$

holds for any Gaussian measure μ on a separable Banach space F and any Borel sets A and B in F (cf. [5]). However the log-concavity of the measure does not imply the Gaussian isoperimetry.

In the early 80's A. Ehrhard [9] gave a different proof of the isoperimetric inequality (2.1) using a Gaussian symmetrization procedure similar to the Steiner symmetrization. With the same symmetrization tool Ehrhard established a new Brunn-Minkowski type inequality, stronger than (3.1), however only for convex sets.

Theorem 3.1(Ehrhard's inequality) *If μ is a centered Gaussian measure on a separable Banach space F and A, B are Borel sets in F, with at least one of them convex, then*

$$\Phi^{-1}(\mu(\lambda A + (1-\lambda)B)) \geq \lambda\Phi^{-1}(\mu(A)) + (1-\lambda)\Phi^{-1}(\mu(B)) \text{ for } \lambda \in [0,1]. \tag{3.2}$$

For both sets A and B convex Ehrhard's inequality was proved in [9]. The generalization to the case when only one of the sets is convex was established in [16].

It is not hard to see that Theorem 3.1 implies the isoperimetric inequality (2.1). Indeed we have for any Borel set A in $\mathbb{R}^n$

$$\Phi^{-1}(\gamma_n(A_t)) = \Phi^{-1}(\gamma_n(\lambda(\lambda^{-1}A) + (1-\lambda)((1-\lambda)^{-1}tB_2^n)))$$
$$\geq \lambda\Phi^{-1}(\gamma_n(\lambda^{-1}A)) + (1-\lambda)\Phi^{-1}(\gamma_n((1-\lambda)^{-1}tB_2^n)) \stackrel{\lambda\to 1-}{\longrightarrow} \Phi^{-1}(\gamma_n(A)) + t.$$

Conjecture 3.1 *Inequality (3.2) holds for any Borel sets in F.*

Ehrhard's symmetrization procedure enables us to reduce Conjecture 3.1 to the case $F = \mathbb{R}$ and $\mu = \gamma_1$. We may also assume that A and B are finite unions of intervals. At the moment the conjecture is known to hold when A is a union of at most 3 intervals.

Ehrhard's inequality has the following Prekopa-Leindler type functional version. Suppose that $\lambda \in (0,1)$ and $f, g, h : \mathbb{R}^n \to [0,1]$ are such that

$$\forall_{x,y\in\mathbb{R}^n}\ \Phi^{-1}(h(\lambda x + (1-\lambda)y)) \geq \lambda\Phi^{-1}(f(x)) + (1-\lambda)\Phi^{-1}(g(y))$$

then

$$\Phi^{-1}(\int_{\mathbb{R}^n} h d\gamma_n) \geq \lambda\Phi^{-1}(\int_{\mathbb{R}^n} f d\gamma_n) + (1-\lambda)\Phi^{-1}(\int_{\mathbb{R}^n} g d\gamma_n). \tag{3.3}$$

We use here the convention $\Phi^{-1}(0) = -\infty, \Phi^{-1}(1) = \infty$ and $-\infty + \infty = -\infty$. At the moment the above functional inequality is known to hold under the additional assumption that at least one of the functions $\Phi^{-1}(f), \Phi^{-1}(g)$ is convex. When one takes $f = 1_A$, $g = 1_B$ and $h = 1_{\lambda A+(1-\lambda)B}$ the inequality (3.3) immediately implies (3.2). On the other hand if we put $A = \{(x,y) \in \mathbb{R}^n \times \mathbb{R} : y \leq \Phi^{-1}(f(x))\}$ and $B = \{(x,y) \in \mathbb{R}^n \times \mathbb{R} : y \leq \Phi^{-1}(g(x))\}$ then $\lambda A + (1-\lambda)B \subset \{(x,y) \in \mathbb{R}^n \times \mathbb{R} : y \leq \Phi^{-1}(h(x))\}$, so Ehrhard's inequality in $\mathbb{R}^{n+1}$ implies (3.3) in $\mathbb{R}^n$. It is easy to show the inductive step in the proof of (3.3). Unfortunately the case $n = 1$ in the functional inequality seems to be much more complicated than the case $\mu = \gamma_1$ in Ehrhard's inequality.

4. Bobkov's inequality

Isoperimetric inequality for the Lebesgue measure has an equivalent analytic form - the Sobolev inequality (cf. [25]). L. Gross [10] showed that the Gaussian measures γ_n satisfy the logarithmic Sobolev inequality

$$\int_{\mathbb{R}^n} g^2 \log g^2 d\gamma_n - \int_{\mathbb{R}^n} g^2 d\gamma_n \log(\int_{\mathbb{R}^n} g^2 d\gamma_n) \leq 2\int_{\mathbb{R}^n} |\nabla g|^2 d\gamma_n \tag{4.1}$$

for all smooth functions $g : \mathbb{R}^n \to \mathbb{R}$. Using the so-called Herbst argument one can show (cf. [19, Sect. 5.1]) that (4.1) implies the concentration inequality

$$\gamma_n(\{h \geq \int_{\mathbb{R}^n} h d\gamma_n + t\}) \leq e^{-t^2/2},\ t \geq 0$$

valid for all Lipschitz functions $h : \mathbb{R}^n \to \mathbb{R}$ with the Lipschitz seminorm $\|h\|_{\mathrm{Lip}} = \sup\{|h(x) - h(y)| : x, y \in \mathbb{R}^n\} \leq 1$. However the logarithmic Sobolev inequality does not imply the isoperimetric inequality.

The formulation of the functional form of Gaussian isoperimetry was given by S.G. Bobkov [2].

Theorem 4.1 *For any locally Lipschitz function* $f : \mathbb{R}^n \to [0,1]$ *and* $\mu = \gamma_n$ *we have*

$$I(\int_{\mathbb{R}^n} f d\mu) \leq \int_{\mathbb{R}^n} \sqrt{I(f)^2 + |\nabla f|^2} d\mu. \tag{4.2}$$

Theorem 4.1 easily implies the isoperimetric inequality (2.2) by approximating the indicator function I_A by Lipschitz functions. On the other hand if we apply (2.2) to the set $A = \{(x,y) \in \mathbb{R}^n \times \mathbb{R} : \Phi(y) < f(x)\}$ in $\mathbb{R}^{n+1}$ we get (4.2). It is also not hard to derive the logarithmic Sobolev inequality (4.1) as a limit case of Bobkov's inequality (cf. [1]): one should use (4.2) for $f = \varepsilon g^2$ (with g bounded) and let ε tend to 0 ($I(t) \sim t\sqrt{2\log(1/t)}$ as $t \to 0+$).

The crucial point of the inequality (4.2) is its tensorization property. To state it precisely let us say that a measure μ on $\mathbb{R}^n$ *satisfies Bobkov's inequality* if the inequality (4.2) holds for all locally Lipschitz functions $f : \mathbb{R}^n \to [0,1]$. Easy argument shows that if μ_i are measures on $\mathbb{R}^{n_i}$, $i = 1,2$, that satisfy Bobkov's inequality then the measure $\mu_1 \otimes \mu_2$ also satisfies Bobkov's inequality.

The inequality (4.2) was proved by Bobkov in an elementary way, based on the following "two-point" inequality:

$$I(\frac{a+b}{2}) \leq \frac{1}{2}\sqrt{I(a)^2 + (\frac{a-b}{2})^2} + \frac{1}{2}\sqrt{I(b)^2 + (\frac{a-b}{2})^2} \tag{4.3}$$

valid for all $a, b \in [0,1]$. In fact the inequality (4.3) is equivalent to Bobkov's inequality for $\mu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ and the discrete gradient instead of ∇f. Using the tensorization property and the central limit theorem Bobkov deduces (in the similar way as Gross in his proof of (4.1)) (4.2) from (4.3).

Using the co-area formula and Theorem 4.1 F. Barthe and M. Maurey [1] gave interesting characterization of all absolutely continuous measures that satisfy Bobkov's inequality.

Theorem 4.2 *Let* $c > 0$ *and* μ *be a Borel probability measure on the Riemannian manifold* M*, absolutely continuous with respect to the Riemannian volume. Then the following properties are equivalent*
(i) For every measurable $A \subset M$, $\mu^+(A) \geq cI(\mu(A))$*;*
(ii) For every locally Lipschitz function $f : M \to [0,1]$

$$I(\int_M f d\mu) \leq \int_M \sqrt{I(f)^2 + \frac{1}{c^2}|\nabla f|^2} d\mu.$$

Theorem 4.2 together with the tensorization property shows that if $\mathrm{Is}(\mu_i) \geq cI$, $i = 1,2\ldots$, then also $\mathrm{Is}(\mu_1 \otimes \ldots \otimes \mu_n) \geq cI$. In general it is not known how to estimate $\mathrm{Is}(\mu_1 \otimes \ldots \otimes \mu_n)$ in terms of $\mathrm{Is}(\mu_i)$ even in the case when all μ_i's are equal (another important special case of this problem was solved in [3]) .

5. S-inequality

In many problems arising in probability in Banach spaces one needs to estimate the measure of balls in some Banach space F. In particular one may ask what is the slowest possible grow of the Gaussian measure of balls in F or more general of some fixed convex symmetric closed set under dilations. The next theorem, proved by R. Latała and K. Oleszkiewicz [17], gives the positive answer to the conjecture posed in an unpublished manuscript of L. A. Shepp (1969).

Theorem 5.1(S-inequality) *Let μ be a centered Gaussian measure on a separable Banach space F. If A is a symmetric, convex, closed subset of F and $P \subset F$ is a symmetric strip, that is $P = \{x \in F : |x^*| \leq 1\}$ for some $x^* \in F^*$, such that $\mu(A) = \mu(P)$ then*

$$\mu(tA) \geq \mu(tP) \text{ for } t \geq 1$$

and

$$\mu(tA) \leq \mu(tP) \text{ for } 0 \leq t \leq 1.$$

A simple approximation argument shows that it is enough to prove Theorem 5.1 for $F = \mathbb{R}^n$ and $\mu = \gamma_n$. The case $n \leq 3$ was solved by V.N. Sudakov and V.A. Zalgaller [30]. Under the additional assumptions of symmetry of A in $\mathbb{R}^n$ with respect to each coordinate, Theorem 5.1 was proved by S. Kwapień and J. Sawa [15].

S-inequality can be equivalently expressed as

$$\Psi^{-1}(\mu(tA)) \geq t\Psi^{-1}(\mu(A)) \text{ for } t \geq 1,$$

where Ψ^{-1} denotes the inverse of

$$\Psi(x) = \gamma_1(-x, x) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} e^{-y^2/2} dy.$$

The crucial tool in the proof of S-inequality is the new modified isoperimetric inequality. Let us first define for a convex symmetric set A in $\mathbb{R}^n$

$$w(A) = 2\sup\{r : B(0,r) \subset A\}.$$

It is easy to see that for a symmetric strip P, $w(P)$ is equal to the width of P and for a symmetric convex set A

$$w(A) = \inf\{w(P) : A \subset P, P \text{ is a symmetric strip in } \mathbb{R}^n\}. \tag{5.1}$$

Thus $w(A)$ can be considered as the width of the set A. The following isoperimetric-type theorem holds true.

Theorem 5.2 *If $\gamma_n(A) = \gamma_n(P)$, where P is a symmetric strip and A is a convex symmetric set in $\mathbb{R}^n$, then*

$$w(A)\gamma_n^+(A) \geq w(P)\gamma_n^+(P). \tag{5.2}$$

The main advantage of the inequality (5.2) is that one may apply here the symmetrization procedure and reduce Theorem 5.2 to the similar statement for 2-dimensional convex sets symmetric with respect to some axis.

It is not hard to see that Theorem 5.2 implies Theorem 5.1. Indeed, let us define for any measurable set B in $\mathbb{R}^n$, $\gamma_B(t) = \gamma_n(tB)$ for $t > 0$. Taking the derivatives of both sides of the inequalities in Theorem 5.1 one can see that it is enough to show

$$\gamma_n(A) = \gamma_n(P) \Rightarrow \gamma'_A(1) \geq \gamma'_P(1) \tag{5.3}$$

for any symmetric convex closed set A and a symmetric strip $P = \{|x_1| \leq p\}$. Let $w = w(A)$, so $B(0,w) \subset A$. Then for $t > 1$ and $x \in A$ we have $B(t^{-1}x, (t-1)w/t) = t^{-1}x + (1-t^{-1})B(0,w) \subset A$, so $B(x,(t-1)w) \subset tA$. Hence $A_{(t-1)w} \subset tA$ and

$$\gamma'_A(1) \geq w\gamma_n^+(A) = w(A)\gamma_n^+(A).$$

However for the strip P

$$\gamma'_P(1) = \sqrt{\frac{2}{\pi}} p e^{-p^2/2} = w(P)\gamma_n^+(P)$$

and the inequality (5.3) follows by Theorem 5.2.

It is not clear if the convexity assumption for the set A in Theorem 5.2 is necessary (obviously $w(A)$ for nonconvex symmetric sets A should be defined by (5.1)). One may also ask if the symmetry assumption can be released (with the suitable modification of the definition of the width for nonsymmetric sets). Also functional versions of Theorems 5.1 and 5.2 are not known.

As was noticed by S. Szarek S-inequality implies the best constants in comparison of moments of Gaussian vectors (cf. [17]).

Corollary 5.3 *If X is a centered Gaussian vector in a separable Banach space $(F, \|\cdot\|)$ then*

$$(\mathbf{E}\|X\|^p)^{1/p} \leq \frac{c_p}{c_q}(\mathbf{E}\|X\|^q)^{1/q} \text{ for any } p \geq q \geq 0,$$

where

$$c_p = (\mathbf{E}|g_1|^p)^{1/p} = \sqrt{2}\left(\frac{1}{\sqrt{\pi}}\Gamma\left(\frac{p+1}{2}\right)\right)^{1/p}.$$

Another interesting problem connected with the S-inequality was recently posed by W. Banaszczyk (private communication): Is it true that under the assumptions of Theorem 5.1

$$\mu(s^\lambda t^{1-\lambda} A) \geq \mu(sA)^\lambda \mu(tA)^{1-\lambda}, \ \lambda \in [0,1] \tag{5.4}$$

for any closed convex symmetric set A in F and $s, t > 0$? Combining the facts that the function $\Phi^{-1}(\mu(tA))$ is concave (Theorem 3.1) and the function $\frac{1}{t}\Psi^{-1}(\mu(tA))$ is nondecreasing (Theorem 5.1) one can show that (5.4) holds if $\mu(sA), \mu(tA) \geq c$, where $c < 0.85$ is some absolute constant.

It is of interest if Theorem 5.1 can be extended to the more general class of measures. The following conjecture seems reasonable.

Conjecture 5.1 *Let ν be a rotationally invariant measure on $\mathbb{R}^n$, absolutely continuous with respect to the Lebesgue measure with the density of the form $f(|x|)$*

for some nondecreasing function $f : \mathbb{R}_+ \to [0, \infty)$. *Then for any convex symmetric set* A *in* $\mathbb{R}^n$ *and any symmetric strip* P *in* $\mathbb{R}^n$ *such that* $\nu(A) = \nu(P)$ *the inequality* $\nu(\lambda A) \geq \nu(\lambda P)$ *is satisfied for* $\lambda \geq 1$.

To show Conjecture 5.1 it is enough to establish the following conjecture concerning the volumes of the convex hulls of symmetric sets on the $n-1$-dimensional unit sphere S^{n-1}.

Conjecture 5.2 *Let* σ_{n-1} *be a Haar measure on* S^{n-1}, A *be a symmetric subset of* S^{n-1} *and* $P = \{x \in S^{n-1} : |x_1| \leq t\}$ *be a symmetric strip on* S^{n-1} *such that* $\sigma_{n-1}(A) = \sigma_{n-1}(P)$, *then* $\mathrm{vol}_n(\mathrm{conv}(A)) \geq \mathrm{vol}_n(\mathrm{conv}(P))$.

It is known that both conjectures hold for $n \leq 3$ (cf. [30]).

6. Correlation conjecture

The following conjecture is an object of intensive efforts of many probabilists since more then 30 years.

Conjecture 6.1 *If* μ *is a centered Gaussian measure on a separable Banach space* F *then*

$$\mu(A \cap B) \geq \mu(A)\mu(B) \tag{6.1}$$

for all convex symmetric sets A, B *in* F.

Various equivalent formulations of Conjecture 6.1 and history of the problem can be found in [27]. Standard approximation argument shows that it is enough to show (6.1) for $F = \mathbb{R}^n$ and $\mu = \gamma_n$. For $n = 2$ the solution was given by L. Pitt [26], for $n \geq 3$ the conjecture remains unsettled, but a variety of special results are known. Borell [8] established (6.1) for sets A, B in a certain class of (not necessary convex) sets in $\mathbb{R}^n$, which for $n = 2$ includes all symmetric sets. A special case of (6.1), when one of the sets A, B is a symmetric strip of the form $\{x \in F : |x^*(x)| \leq 1\}$ for some $x^* \in F^*$, was proved independently by C. G. Khatri [14] and Z. Šidák [28] (see [11] for an extension to elliptically contoured distributions and [31] for the case when one of the sets is a nonsymmetric strip). Recently, the Khatri-Šidák result has been generalized by G. Hargé [12] to the case when one of the sets is a symmetric ellipsoid.

Theorem 6.1 *If* μ *is a centered Gaussian measure on* $\mathbb{R}^n$, A *is a symmetric convex set in* $\mathbb{R}^n$ *and* B *is a symmetric ellipsoid, that is the set of the form* $B = \{x \in \mathbb{R}^n : \langle Cx, x\rangle \leq 1\}$ *for some symmetric nonnegative matrix* C, *then*

$$\mu(A \cap B) \geq \mu(A)\mu(B).$$

The following weaker form of (6.1)

$$\mu(A \cap B) \geq \mu(\lambda A)\mu(\sqrt{1-\lambda^2}B), \ 0 \leq \lambda \leq 1$$

was established for $\lambda = \frac{1}{\sqrt{2}}$ in [27] and for general λ in [21]. The Khatri-Šidák result and the above inequality turn out to be very useful in the study of the so-called small ball probabilities for Gaussian processes (see [22] for a survey of results in this direction).

The correlation conjecture has the following functional form:

$$\int f g d\mu \geq \int f d\mu \int g d\mu \tag{6.2}$$

for all nonnegative even functions f, g such that the sets $\{f \geq t\}$ and $\{g \geq t\}$ are convex for all $t \geq 0$. Y. Hu [13] showed that the inequality (6.2) (that we would like to have for log-concave functions) is valid for even convex functions $f, g \in L^2(F, \mu)$.

References

[1] F. Barthe, B. Maurey, Some remarks on isoperimetry of Gaussian type, *Ann. Inst. H. Poincaré Probab. Statist.* 36 (2000), 419–434.

[2] S.G. Bobkov, An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in Gauss space, *Ann. Probab.* 25 (1997), 206–214.

[3] S.G. Bobkov, C. Houdré, Isoperimetric constants for product probability measures, *Ann. Probab.* 25 (1997), 184–205.

[4] V.I. Bogachev, *Gaussian Measures*, American Mathematical Society, Providence, RI, 1998.

[5] C. Borell, Convex measures on locally convex spaces, *Ark. Mat.* 12 (1974), 239–252.

[6] C. Borell, The Brunn-Minkowski inequality in Gauss space, *Invent. Math.*, 30 (1975), 207–216.

[7] C. Borell, Gaussian Radon measures on locally convex spaces, *Math. Scand.* 38 (1976), 265–284.

[8] C. Borell, A Gaussian correlation inequality for certain bodies in R^n, *Math. Ann.* 256 (1981), 569–573.

[9] A. Ehrhard, Symétrisation dans l'espace de Gauss, *Math. Scand.*, 53 (1983), 281–301.

[10] L. Gross, Logaritmic Sobolev inequalities, *Amer. J. Math.* 97 (1975), 1061–1083.

[11] S. Das Gupta, M.L. Eaton, I. Olkin, M. Perlman, L.J. Savage, M. Sobel, Inequalities on the probability content of convex regions for elliptically contoured distributions, *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* vol. II, 241–264, Univ. California Press, Berkeley, 1972.

[12] G. Hargé, A particular case of correlation inequality for the Gaussian measure, *Ann. Probab.* 27 (1999), 1939–1951.

[13] Y. Hu, Itô-Wiener chaos expansion with exact residual and correlation, variance inequalities, *J. Theoret. Probab.* 10 (1997), 835–848.

[14] C.G. Khatri, On certain inequalities for normal distributions and their applications to simultaneous confidence bounds, *Ann. Math. Stat.* 38 (1967), 1853–1867.

[15] S. Kwapień, J. Sawa, On some conjecture concerning Gaussian measures of dilatations of convex symmetric sets, *Studia Math.* 105 (1993), 173–187.

[16] R. Latała, A note on the Ehrhard inequality, *Studia Math.* 118 (1996), 169–174.

[17] R. Latała, K. Oleszkiewicz, Gaussian measures of dilatations of convex symmetric sets, *Ann. Probab.* 27 (1999), 1922–1938.
[18] M. Ledoux, Isoperimetry and Gaussian Analysis, *Lectures on probability theory and statistics (Saint-Flour, 1994)*, 165–294, Lecture Notes in Math. 1648, Springer, Berlin, 1996.
[19] M. Ledoux, *The concentration of measure phenomenon*, American Mathematical Society, Providence, RI, 2001.
[20] M. Ledoux, M. Talagrand, *Probability on Banach Spaces. Isoperimetry and processes*, Springer-Verlag, Berlin, 1991.
[21] W.V. Li, A Gaussian correlation inequality and its applications to small ball probabilities, *Electron. Comm. Probab.* 4 (1999), 111-118.
[22] W.V. Li, Q.M. Shao, Gaussian processes: inequalities, small ball probabilities and applications, *Stochastic Processes: Theory and Methods*, Handbook of Statistics vol. 19, 533–597, Elsevier, Amsterdam 2001.
[23] M.A. Lifshits, *Gaussian random functions*, Kluwer Academic Publications, Dordrecht, 1995.
[24] V.D. Milman, G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces*, Lecture Notes in Math. 1200, Springer-Verlag, Berlin, 1986.
[25] R. Osserman, The isoperimetric inequality, *Bull. Amer. Math. Soc.* 84 (1978), 1182–1238.
[26] L. Pitt, A Gaussian correlation inequality for symmetric convex sets, *Ann. Probability*, 5 (1977), 470–474.
[27] G. Schechtman, T. Schlumprecht, J. Zinn, On the Gaussian measure of intersection, *Ann. Probab.* 26 (1998), 346–357.
[28] Z. Šidák, Rectangular confidence regions for the means of multivariate normal distributions, *J. Amer. Statist. Assoc.* 62 (1967), 626–633.
[29] V.N. Sudakov, B.S. Tsirel'son, Extremal propertiesof half-spaces for spherically invariant measures (in Russian), *Zap. Nauchn. Sem. L.O.M.I.* 41 (1974), 14–24.
[30] V.N. Sudakov, V.A. Zalgaller, Some problems on centrally symmetric convex bodies (in Russian) *Zap. Nauchn. Sem. L.O.M.I.* 45 (1974), 75–82.
[31] S. Szarek, E. Werner, A nonsymmetric correlation inequality for Gaussian measure, *J. Multivariate Anal.* 68 (1999), 193–211.
[32] M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces, *IHES Publ. Math.* 81 (1995), 73–205.

Author Index

Alesker, Semyon 757
Andrews, B. 221
Bartnik, Robert 231
Bestvina, Mladen 373
Biane, P. 765
Bigelow, S. 37
Biran, P. 241
Bisch, D. 775
Bondal, A. 47
Bouscaren, E. 3
Bray, Hubert L. 257
Chekanov, Yu. V. 385
Chen, Xiuxiong 273
Cogdell, J. W. 119
Cohen, H. 129
Delorme, Patrick 545
Denef, J. 13
Ding, Weiyue 283
Eremenko, A. 681
Esnault, Hélène 471
Etingof, Pavel 555
Fontaine, Jean-Marc 139
Furuta, M. 395
Gaitsgory, D. 571
Ge, Liming 787
Giroux, Emmanuel 405
Göttsche, L. 483
Harris, Michael 583
Heinonen, Juha 691
Hesselholt, Lars 415
Huber, A. 149
Ionel, Eleny-Nicoleta 427
Kato, Kazuya 163
Kenig, Carlos E. 701
Kings, G. 149
Klyachko, Alexander 599
Kobayashi, Toshiyuki 615
Kudla, Stephen S. 173
Lafforgue, V. 795
Lascar, D. 25
Latała, R. 813
Lerner, Nicolas 711
Levine, M. 57
Li, P. 293
Loeser, F. 13
Long, Yiming 303
Mazur, Barry 185
Mehta, Vikram Bhagvandas ..629
Meinrenken, E. 637
Mukai, Shigeru 495
Nazarov, Maxim 643
Orlov, D. 47
Pandharipande, R. 503
Petrunin, Anton 315
Piatetski-Shapiro, I. I. 119
Praeger, Cheryl E. 67
Reid, Miles 513
Rong, Xiaochun 323
Rost, Markus 77
Rubin, Karl 185
Schechtman, Vadim 525
Schwartz, Richard Evan 339
Seidel, Paul 351
Sela, Z. 87
Shahidi, Freydoon 655
Stafford, J. T. 93
Tamarkin, Dimitri 105
Teichner, Peter 437
Thiele, C. 721
Tillmann, Ulrike 447
Totaro, B. 533
Ullmo, Emmanuel 197
Vignéras, Marie-France 667
Wang, Shicheng 457
Wooley, Trevor D. 207
Zelditch, S. 733
Zhang, Weiping 361
Zhou, Xiangyu 743