Contents

1 Logic and Foundations

2 Algebra

In case of papers with several authors, invited speakers at the Congress are marked with an asterisk.

3 Number Theory

4 Algebraic and Complex Geometry

5 Geometry

6 Topology

7 Lie Theory and Generalizations

8 Analysis

9 Functional Analysis and Applications

10 Dynamical Systems and Ordinary Differential Equations

11 Partial Differential Equations

12 Mathematical Physics

13 Probability and Statistics

14 Combinatorics

15 Mathematical Aspects of Computer Science

16 Numerical Analysis and Scientific Computing

17 Control Theory and Optimization

18 Mathematics in Science and Technology

19 Mathematics Education and Popularization of Mathematics

20 History of Mathematics

Section 7

Lie Theory and Generalizations

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Quasi-isometric Rigidity of Solvable Groups

Alex Eskin* and David Fisher†

Abstract

In this article we survey recent progress on quasi-isometric rigidity of polycyclic groups. These results are contributions to Gromov's program for classifying finitely generated groups up to quasi-isometry [Gr2]. The results discussed here rely on a new technique for studying quasi-isometries of finitely generated groups, which we refer to as *coarse differentiation*.

We include a discussion of other applications of coarse differentiation to problems in geometric group theory and a comparison of coarse differentiation to other related techniques in nearby areas of mathematics.

Mathematics Subject Classification (2010). Primary 22E25; Secondary 20F65.

Keywords. Quasi-isometry, rigidity, polycyclic groups.

1. Introduction, Conjectures, and Results

For any group Γ generated by a subset S one has the associated Cayley graph, $C_\Gamma(S)$. This is the graph with vertex set Γ and edges connecting any pair of elements which differ by right multiplication by a generator. There is a natural Γ action on $C_\Gamma(S)$ by left translation. By giving every edge length one, the Cayley graph can be made into a (geodesic) metric space. The distance on Γ viewed as the vertices of the Cayley graph is the *word metric*, defined via the norm:

$$\|\gamma\| = \inf\{\text{length of a word in the generators } S \text{ representing } \gamma \text{ in } \Gamma.\}$$

First author partially supported by NSF grant DMS-0905912. Second author partially supported by NSF grant DMS-0643546.

*Department of Mathematics, University of Chicago, 5734 S. University Avenue, Chicago, Illinois 60637.

†Department of Mathematics, Indiana University, Rawles Hall, Bloomington, IN, 47405.

Different sets of generators give rise to different metrics and Cayley graphs for a group but one wants these to be equivalent. The natural notion of equivalence in this category is *quasi-isometry*:

Definition 1.1. *Let* (X, d_X) *and* (Y, d_Y) *be metric spaces. Given real numbers* $K \geq 1$ *and* $C \geq 0$, *a map* $f : X \to Y$ *is called a* (K, C)-quasi-isometry *if*

1. $\frac{1}{K} d_X(x_1, x_2) - C \leq d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2) + C$ *for all* x_1 *and* x_2 *in* X, *and,*
2. *the* C *neighborhood of* $f(X)$ *is all of* Y.

If Γ is a finitely generated group, Γ is canonically quasi-isometric to any finite index subgroup Γ' in Γ and to any quotient $\Gamma'' = \Gamma/F$ for any finite normal subgroup F. The equivalence relation generated by these (trivial) quasi-isometries is called *weak commensurability*. A group is said to *virtually* have a property if some weakly commensurable group does.

In his ICM address in 1983, Gromov proposed a broad program for studying finitely generated groups as geometric objects, [Gr2]. Though there are many aspects to this program (see [Gr3] for a discussion), the principal question is the classification of finitely generated groups up to quasi-isometry. By construction, any finitely generated group Γ is quasi-isometric to any space on which Γ acts properly discontinuously and cocompactly by isometries. For example, the fundamental group of a compact manifold is quasi-isometric to the universal cover of the manifold (this is called the Milnor–Svarc lemma). In particular, any two cocompact lattices in the same Lie group G are quasi-isometric. One important aspect of Gromov's program is that it allows one to generalize many invariants, techniques, and questions from the study of lattices to all finitely generated groups.

A major direction in the Gromov program is determining which algebraic properties of groups are quasi-isometry invariants. As consequence of Gromov's theorem on groups of polynomial growth, one has that the property of having a finite index subgroup that is nilpotent is invariant under quasi-isometries [Gr1]. It is then an obvious question whether larger classes of groups might have this property. Erschler showed in [D] that this is not the case for solvable groups. I.e. there are groups quasi-isometric to solvable groups which are not even virtually solvable. However, the following conjecture is plausible and we will spend much of this article discussing progress towards it.

Conjecture 1.2. *Let* Γ *be a polycyclic group, then any group* Γ' *quasi-isometric to* Γ *is virtually polycyclic.*

Remarks:

1. Conjecture 1.2 can be rephrased as being about lattices in connected, simply connected solvable Lie groups. In particular, by a theorem of Mostow, any polycyclic group is virtually a lattice in a connected, simply connected

solvable Lie group, and conversely any lattice in a solvable Lie group is virtually polycyclic [Mo2]. As solvable Lie groups have only cocompact lattices, the conjectures is equivalent to saying that any group quasi-isometric to lattice in a simply connected solvable Lie group is virtually a lattice in a simply connected, solvable Lie group.

2. Some classes of solvable groups which are not polycyclic are known to be quasi-isometrically rigid. See particularly the work of Farb and Mosher on the solvable Baumslag–Solitar groups [FM1, FM2] as well as later work of Farb–Mosher, Mosher–Sageev–Whyte and Wortman [FM3, MSW, W]. The methods used in all of these works depend essentially on topological arguments based on the explicit structure of singularities of the spaces studied and cannot apply to polycyclic groups.

3. Shalom has obtained some evidence for the conjecture by cohomological methods [Sh]. For example, Shalom shows that any group quasi-isometric to a polycyclic group has a finite index subgroup with infinite abelianization. Some of his results have been further refined by Sauer [Sa].

We discuss results that establish Conjecture 1.2 in many cases. We believe our techniques provide a method to attack the conjecture. This is work in progress, joint with Irine Peng.

From an algebraic point of view, solvable groups are generally easier to study than semisimple ones, as the algebraic structure is more easily manipulated. In the present context it is extremely difficult to see that any algebraic structure is preserved and so we are forced to work geometrically. For nilpotent groups the only geometric fact needed is polynomial volume growth. For semisimple groups, the key fact for all approaches is nonpositive curvature. The geometry of solvable groups is quite difficult to manage, since it involves a mixture of positive and negative curvature as well as exponential volume growth.

The simplest non-trivial example for Conjecture 1.2 is the 3-dimensional solvable Lie group Sol. This example has received a great deal of attention. The group Sol $\cong \mathbb{R}\ltimes\mathbb{R}^2$ with $\mathbb{R}$ acting on $\mathbb{R}^2$ via the diagonal matrix with entries $e^{z/2}$ and $e^{-z/2}$. As matrices, Sol can be written as :

$$\text{Sol} = \left\{ \left(\begin{array}{ccc} e^{z/2} & x & 0 \\ 0 & 1 & 0 \\ 0 & y & e^{-z/2} \end{array} \right) \middle| \, (x,y,z) \in \mathbb{R}^3 \right\}$$

The metric $e^{-z}dx^2 + e^z dy^2 + dz^2$ is a left invariant metric on Sol. Any group of the form $\mathbb{Z}\ltimes_T\mathbb{Z}^2$ for $T \in SL(2,\mathbb{Z})$ with $|tr(T)| > 2$ is a cocompact lattice in Sol.

The following theorem by Eskin, Fisher and Whyte proves a conjecture of Farb and Mosher [EFW0, EFW1, EFW2, FM4]:

Theorem 1.3. *Let* Γ *be a finitely generated group quasi-isometric to* Sol. *Then* Γ *is virtually a lattice in* Sol.

Peng's thesis contains a far reaching generalization of this result [Pe1, Pe2]. In addition to generalizing the methods introduced in [EFW0, EFW1, EFW2], Peng's thesis makes use of generalizations of some results of Farb and Mosher by Dymarz and Dymarz–Peng [FM1, Dy, DP]. We require some vocabulary to formulate Peng's results. We call a solvable Lie group *abelian by abelian* if it is of the form $\mathbb{R}^k \ltimes \mathbb{R}^n$. Such a group is defined by a linear represenation $\rho : \mathbb{R}^k \to GL(\mathbb{R}^n)$. Note that the image $\rho(\mathbb{R}^k)$ is an abelian subgroup of $GL(\mathbb{R}^n)$ and as such it's elements admit a common Jordan form. The Jordan form gives rise to a collection of functionals, called *weights*, on $\mathbb{R}^k$. Each weight ω corresponds to a subspace W of $\mathbb{R}^n$ that is common generalized eigenspace for the $\mathbb{R}^k$ action and $\omega(v)$ for v in $\mathbb{R}^k$ is the norm of the generalized eigenvalue for the action of v on W. We call an abelian by abelian solvable group non-degenerate if ρ is faithful and no weight w has $w(\mathbb{R}^k)$ contained in $\{\pm 1\}$. Recall that a Lie group is unimodular if it has a bi-invariant Haar measure. For a group of the form $\mathbb{R}^k \ltimes \mathbb{R}^n$, unimodularity is equivalent to ρ taking values in $SL(\mathbb{R}^n)$.

Theorem 1.4. *Let $G = \mathbb{R}^k \ltimes \mathbb{R}^n$ be a non-degenerate abelian by abelian unimodular, solvable Lie group. Then any group Γ quasi-isometric to G is virtually a lattice in a solvable Lie group $G' = \mathbb{R}^k \ltimes \mathbb{R}^n$ which is also abelian by abelian, non-degenerate and unimodular.*

Remark: One can in fact say more about the relation between G and G', but we will not pursue this here.

Both of the theorems stated above are proved using a new technique, which we call *coarse differentiation*. Even though quasi-isometries have no local structure and conventional derivatives do not make sense, we essentially construct a "coarse derivative" that models the large scale behavior of the quasi-isometry. Coarse differentiation is quite similar to a number of notions that arise in various forms of differentiation theory. However, this construction is quite different from the more conventional method of passing to the asymptotic cone and then applying a differentiation theorem to either the full asymptotic cone or some subspace of it, see §4.5 for more discussion.

2. Quasi-isometries are Height Respecting

A typical step in the study of quasi-isometric rigidity of groups is the identification of all quasi-isometries of some space X quasi-isometric to the group, see §4.6 for a brief explanation. For us, the space X is always a solvable Lie group. To pursue Conjecture 1.2, the goal is to show that all self quasi-isometries of the solvable Lie group G permutes the cosets of a certain subgroup.

For Sol the group whose cosets we show are preserved is exactly the kernel of the homomorphism $h : \mathrm{Sol} \to \mathbb{R}$ which we call the height function. There is a foliation of Sol by level sets of the height function which is also the foliation

by cosets of the normal $\mathbb{R}^2$. We will call a quasi-isometry of any of these spaces *height respecting* if it permutes the height level sets to within bounded distance (In [FM4], the term used is horizontal respecting). In our coordinates for Sol, the height function is $h(x, y, z) = z$.

Theorem 2.1. *Any (K, C)-quasi-isometry φ of* Sol *is within bounded distance of a height respecting quasi-isometry $\hat{\varphi}$. Furthermore, this distance can be taken uniform in (K, C) and therefore, in particular, $\hat{\varphi}$ is a (K', C')-quasi-isometry where K', C' depend only on K and C.*

Remark: In fact, Theorem 2.1 can be used to identify the quasi-isometries of Sol completely. Possibly after composing with the map $(x, y, z) \to (y, x, -z)$, any height respecting quasi-isometry (and in particular, any isometry) is at bounded distance from a quasi-isometry of the form $(x, y, z) \to (f(x), g(y), z)$ where f and g are bilipschitz functions. Given a metric space X, one defines $\mathrm{QI}(X)$ to be the group of quasi-isometries of X modulo the subgroup of those at finite distance from the identity. The previous statement can then be taken to mean that $\mathrm{QI}(\mathrm{Sol}) = \mathrm{Bilip}(\mathbb{R})^2 \ltimes \mathbb{Z}/2\mathbb{Z}$. This explicit description was conjectured by Farb and Mosher.

If we take a group of the form $\mathbb{R}^k \ltimes \mathbb{R}^n$ as in Theorem 1.4, we can write coordinates $(z, \vec{x})$ where z is the coordinate in $\mathbb{R}^k$ and $\vec{x}$ is the coordinate in $\mathbb{R}^n$. Here $h(z, \vec{x}) = z$ and level sets of h are $\mathbb{R}^n$ cosets. We can again call a quasi-isometry height respecting if it permutes level sets of h. The following is the main result of [Pe1, Pe2].

Theorem 2.2. *Let $X = \mathbb{R}^k \ltimes \mathbb{R}^n$ be as in Theorem 1.4. Then any (K, C)-quasi-isometry φ of $\mathbb{R}^k \ltimes \mathbb{R}^n$ is within a bounded distance of a height respecting quasi-isometry $\hat{\varphi}$. Furthermore, the bound is uniform in K and C.*

Remark: There is an explicit description of $QI(\mathbb{R}^k \ltimes \mathbb{R}^n)$ in this context as well, but it is somewhat involved so we omit it.

We now describe a conjecture that is a key step in our approach to Conjecture 1.2. We note here that this conjecture does not suffice to prove that one, but that one requires in addition generalizations of the results in [Dy, DP].

We begin by reviewing some structure theory of simply connected solvable Lie groups. Most of the basics are contained in work of Auslander [A1, A2]. Let G be a solvable Lie group. Then G is part of a short exact sequence:

$$1 \to N \to G \to \bar{H} \to 1$$

where N and $\bar{H}$ are nilpotent. In general this exact sequence does not split but there is a nilpotent group H, called a Cartan subgroup, that is minimal among all groups mapping onto $\bar{H}$. The group N is the maximal normal nilpotent subgroup of G, also known as it's nilradical. We are particularly interested in a subgroup $\exp(G)$ of N, defined independently by Guivarch and Osin, called the exponential radical of G [Gu, Os]. This can be taken to be the subgroup of G generated by all exponentially distorted elements in G. (Guivarch calls it

the *unstable* subgroup, the terminology *exponential radical* is due to Osin.) We believe that the following conjecture is a key step for proving Conjecture 1.2.

Conjecture 2.3. *Given a unimodular solvable Lie group G, then any self quasi-isometry of G is at bounded distance from one which preserves the foliation by cosets of* $\exp(G)$.

We remark here that the assumption that G be unimodular is necessary. To see this consider the group $SL(2,\mathbb{R})$. This group is quasi-isometric to the affine group of the line, which is a two dimensional solvable Lie group of the form $\mathrm{Aff}(\mathbb{R}) = \mathbb{R} \ltimes \mathbb{R}$. It is easy to see that the normal $\mathbb{R}$, i.e. the group of translations, is the exponential radical of $\mathrm{Aff}(\mathbb{R})$. However, $\mathrm{Aff}(\mathbb{R})$ is quasi-isometric to $SL(2,\mathbb{R})/O(2) = \mathbb{H}^2$ and the group of quasi-isometries of $\mathbb{H}^2$ is known to be the group of quasi-symmetric maps of $S^1 = \partial\mathbb{H}^2$. The foliation by cosets of $\mathbb{R}$ identifies naturally with the horocyclic foliation corresponding to the fixed point for $\mathrm{Aff}(\mathbb{R}) < SL(2,\mathbb{R})$ on S^1 and it is easy to see that the cosets of this foliation are not permuted by all quasi-symmetric maps.

3. Geometry of Sol

In this subsection we describe the geometry of Sol and related spaces in more detail, with emphasis on the geometric facts used in our proofs.

The upper half plane model of the hyperbolic plane $\mathbb{H}^2$ is the set $\{(x,\xi) \mid \xi > 0\}$ with the length element $ds^2 = \frac{1}{\xi^2}(dx^2 + d\xi^2)$. If we make the change of variable $z = \log\xi$, we get $\mathbb{R}^2$ with the length element $ds^2 = dz^2 + e^{-z}dx^2$. This is the *log model* of the hyperbolic plane $\mathbb{H}^2$.

The length element of Sol is:

$$ds^2 = dz^2 + e^{-z}dx^2 + e^{z}dy^2.$$

Thus planes parallel to the xz plane are hyperbolic planes in the log model. Planes parallel to the yz plane are *upside-down* hyperbolic planes in the log model. All of these copies of $\mathbb{H}^2$ are isometrically embedded and totally geodesic.

We will refer to lines parallel to the x-axis as x-horocycles, and to lines parallel to the y-axis as y-horocycles. This terminology is justified by the fact that each (x or y)-horocycle is indeed a horocycle in the hyperbolic plane which contains it.

We now turn to a discussion of geodesics and quasi-geodesics in Sol. Any geodesic in an $\mathbb{H}^2$ leaf in Sol is a geodesic. There is a special class of geodesics, which we call *vertical geodesics*. These are the geodesics which are of the form $\gamma(t) = (x_0, y_0, t)$ or $\gamma(t) = (x_0, y_0, -t)$. We call the vertical geodesic *upward oriented* in the first case, and *downward oriented* in the second case. In both cases, this is a unit speed parametrization. Each vertical geodesic is a geodesic in two hyperbolic planes, the plane $y = y_0$ and the plane $x = x_0$.

Certain quasi-geodesics in Sol are easy to describe. Given two points (x_0, y_0, t_0) and (x_1, y_1, t_1), there is a geodesic γ_1 in the hyperbolic plane $y = y_0$ that joins (x_0, y_0, t_0) to (x_1, y_0, t_1) and a geodesic γ_2 in the plane $x = x_1$ that joins (x_1, y_0, t_1) to a (x_1, y_1, t_1). It is easy to check that the concatenation of γ_1 and γ_2 is a quasi-geodesic. In first matching the x coordinates and then matching the y coordinates, we made a choice. It is possible to construct a quasi-geodesic by first matching the y coordinates and then the x coordinates. This immediately shows that any pair of points not contained in a hyperbolic plane in Sol can be joined by two distinct quasi-geodesics which are not close together. This is an aspect of positive curvature. One way to prove that the objects just constructed are quasi-geodesics is to note the following: The pair of projections $\pi_1, \pi_2 : \mathrm{Sol} \to \mathbb{H}^2$ onto the xt and yt coordinate planes can be combined into a quasi-isometric embedding $\pi_1 \times \pi_2 : \mathrm{Sol} \to \mathbb{H}^2 \times \mathbb{H}^2$.

We state here the simplest version of a key geometric fact used at various steps in the proof.

Lemma 3.1 (Quadrilaterals). *Suppose $p_1, p_2, q_1, q_2 \in \mathrm{Sol}$ and $\gamma_{ij} : [0, \ell_{ij}] \to \mathrm{Sol}$ are vertical geodesic segments parametrized by arclength. Suppose $C > 0$. Assume that for $i = 1, 2$, $j = 1, 2$,*

$$d(p_i, \gamma_{ij}(0)) \leq C \qquad \text{and} \qquad d(q_i, \gamma_{ij}(\ell_{ij})) \leq C,$$

so that γ_{ij} connects the C-neighborhood of p_i to the C-neighborhood of q_j. Further assume that for $i = 1, 2$ and all t, $d(\gamma_{i1}(t), \gamma_{i2}(t)) \geq (1/10)t - C$ (so that for each i, the two segments leaving the neighborhood of p_i diverge right away) and for $j = 1, 2$ and all t, $d(\gamma_{1j}(l_{1j} - t), \gamma_{2j}(l_{2j} - t)) \geq (1/10)t - C$. Then there exists C_1 depending only on C such that exactly one of the following holds:

(a) *All four γ_{ij} are upward oriented, p_2 is within C_1 of the y-horocycle passing through p_1 and q_2 is within C_1 of the x-horocycle passing through $\phi(q_1)$.*

(b) *All four γ_{ij} are downward oriented, p_2 is within C_1 of the x-horocycle passing through p_1 and q_2 is within C_1 of the y-horocycle passing through q_1.*

We think of p_1, p_2, q_1 and q_2 as defining a quadrilateral. The content of the lemma is that any quadrilateral has its four "corners" in pairs that lie essentially along horocycles. In particular, if we take a quadrilateral with geodesic segments γ_{ij} and with $h(p_1) = h(p_2)$ and $h(q_1) = h(q_2)$ and map it forward under a (K, C)-quasi-isometry ϕ, and if we would somehow know that ϕ sends each of the four γ_{ij} close to a vertical geodesic, then Lemma 3.1 would imply that ϕ sends the p_i (resp. q_i) to a pair of points at roughly the same height.

We now define certain useful subsets of Sol. Let $B(L, \vec{0}) = [-e^L, e^L] \times [-e^L, e^L] \times [-L, L]$. Then $|B(L, \vec{0})| \approx Le^{2L}$ and $Area(\partial B(L, \vec{0})) \approx e^{2L}$, so $B(L)$

is a Fölner set. We call $B(L, \vec{0})$ a box of size L centered at the identity. We define the box of size L centered at a point p by $B(L, p) = T_p B(L, \vec{0})$ where T_p is left translation by p. Since left translation is an isometry, $B(L, p)$ is also a Fölner set. We frequently omit the center of a box in our notation and write $B(L)$.

Approximating a box by a graph. Notice that the top of $B(L)$, meaning the set $[-e^L, e^L] \times [-e^L, e^L] \times \{L\}$, is not at all square - the sides of this rectangle are horocyclic segments of lengths $2e^{2L}$ and 2 - in other words it is just a small metric neighborhood of a horocycle. Similarly, the bottom is also essentially a horocycle but in the transverse direction. Further, we can connect the 1-neighborhood of any point of the top horocycle to the 1-neighborhood of any point of the bottom horocycle by a vertical geodesic segment, and these segments essentially sweep out the box $B(L)$. Thus a box contains an extremely large number of quadrilaterals. If we discretize the top and bottom horocycle, we can think of this process as giving a description of a graph which we call G_L. This graph is essentially a complete bipartite graph with $4e^{2L}$ vertices. Throughout the proof of our results on Sol, this highly connected graph plays a key role.

4. On Proofs

In this section, we give some of the key ideas in the proofs. In the first two subsections we indicate the key new ideas behind our proof of Theorem 2.1. The first contains quantative estimates on the behavior of quasi-geodesics. The second subsection averages this behavior over families of quasi-geodesics. In §4.3 we sketch the proof of Theorem 2.1. Subsection 4.4 briefly discusses the ideas needed to adapt the proof of Theorem 2.1 to prove the other results in Section 2 and indicates obstructions and progress in the general case of Conjecture 2.3. Before continuing with discussion of proofs, we include a discussion of how to axiomatize the methods of §4.1 and §4.2 into a general method of *coarse differentiation* in §4.5. In subsection §4.6, we discuss deducing results in §1 from results in §2.

4.1. Behavior of quasi-geodesics.

We begin by discussing some quantative estimates on the behavior of quasi-geodesic segments in Sol. Throughout the discussion we assume $\alpha : [0, r] \to \text{Sol}$ is a (K, C)-quasi-geodesic segment for a fixed choice of (K, C), i.e. α is a quasi-isometric embedding of $[0, r]$ into Sol. A quasi-isometric embedding is a map that satisfies point (1) in Definition 1.1 but not point (2).

Definition 4.1 (ϵ**-monotone).** *A quasigeodesic segment $\alpha : [0, r] \to \text{Sol}$ is ϵ-monotone if for all $t_1, t_2 \in [0, r]$ with $h(\alpha(t_1)) = h(\alpha(t_2))$ we have $|t_1 - t_2| < \epsilon r$.*

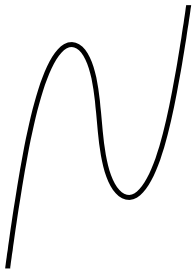


Figure 1. A quasigeodesic segment which is not ϵ-monotone.

The following fact about ε-monotone geodesics is an easy exercise in hyperbolic geometry:

Lemma 4.2 (ϵ-monotone is close to vertical). *If $\alpha : [0, r] \to \mathrm{Sol}$ is ϵ-monotone, then there exists a vertical geodesic segment λ such that $d(\alpha, \lambda) = O(\epsilon r)$.*

Remark: The distance $d(\alpha, \lambda)$ is the Hausdorff distance between the sets and does not depend on parametrizations.

Lemma 4.3 (Subdivision). *Suppose $\alpha : [0, r] \to \mathrm{Sol}$ is a quasi-geodesic segment which is not ϵ-monotone. Suppose $n \gg 1$ (depending on ϵ, K, C). Then*

$$\sum_{j=0}^{n-1} \left| h(\alpha(\tfrac{(j+1)r}{n})) - h(\alpha(\tfrac{jr}{n})) \right| \geq |h(\alpha(0)) - h(\alpha(r))| + \frac{\epsilon r}{8K^2}.$$

Outline of Proof. If n is sufficiently large, the total variation of the height increases after the subdivision by a term proportional to ϵ. See Figure 2. □

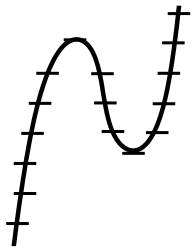


Figure 2. Proof of Lemma 4.3

Choosing Scales: Choose $1 \ll r_0 \ll r_1 \ll \cdots \ll r_M$. In particular, $C \ll r_0$ and $r_{m+1}/r_m > n$.

Lemma 4.4. *Suppose $L \gg r_M$, and suppose $\alpha : [0, L] \to \mathrm{Sol}$ is a quasi-geodesic segment. For each $m \in [1, M]$, subdivide $[0, L]$ into L/r_m segments of*

length r_m. Let $\delta_m(\alpha)$ denote the fraction of these segments whose images are not ϵ-monotone. Then,

$$\sum_{m=1}^{M} \delta_m(\alpha) \leq \frac{16K^3}{\epsilon}.$$

Proof. By applying Lemma 4.3 to each non-ϵ-monotone segment on the scale r_M, we get

$$\sum_{j=1}^{L/r_{M-1}} |h(\alpha(jr_{M-1})) - h(\alpha((j-1)r_{M-1}))| \geq$$
$$\geq \sum_{j=1}^{L/r_M} |h(\alpha(jr_M)) - h(\alpha((j-1)r_M))| + \delta_M(\alpha)\frac{\epsilon L}{8K^2}.$$

Doing this again, we get after M iterations,

$$\sum_{j=1}^{L/r_0} |h(\alpha(jr_0)) - h(\alpha((j-1)r_0))| \geq$$
$$\geq \sum_{j=1}^{L/r_M} |h(\alpha(jr_M)) - h(\alpha((j-1)r_M))| + \frac{\epsilon L}{8K^2}\sum_{m=1}^{M} \delta_m(\alpha).$$

But the left-hand-side is bounded from above by the length and so bounded above by $2KL$. □

4.2. Averaging. In this subsection we apply the estimates from above to images of geodesics under a quasi-isometry of Sol. The idea is to average the previous estimates over families of quasi-geodesics. This results in a coarse analogue of Rademacher's theorem, which says that a bilipschitz map of $\mathbb{R}^n$ is differentiable almost everywhere, see below for discussion.

Setup and Notation.

- Suppose $\phi : \text{Sol} \to \text{Sol}$ is a (K, C) quasi-isometry. Without loss of generality, we may assume that ϕ is continuous.
- Let $\gamma : [-L, L] \to \text{Sol}$ be a vertical geodesic segment parametrized by arclength where $L \gg C$.
- Let $\overline{\gamma} = \phi \circ \gamma$. Then $\overline{\gamma} : [-L, L] \to \text{Sol}$ is a quasi-geodesic segment.

It follows from Lemma 4.4, that for every $\theta > 0$ and every geodesic segment γ, assuming that M is sufficiently large, there exists $m \in [1, M]$ such that

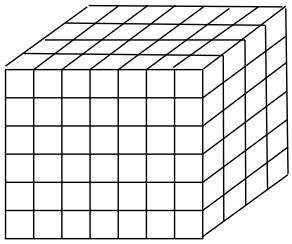


Figure 3. The box $B(L)$.

$\delta_m(\overline{\gamma}) < \theta$. The difficulty is that m may depend on γ. For Sol, this is overcome as follows:

Recall that $B(L) = [-e^L, e^L] \times [-e^L, e^L] \times [-L, L]$. Then $|B(L)| \approx Le^{2L}$ and $Area(\partial B(L)) \approx e^{2L}$, so $B(L)$ is a Fölner set. Average the result of Lemma 4.4 over Y_L, the set of vertical geodesics in $B(L)$ and let $|Y_L|$ denote the measure/cardinality of Y_L. Changing order, we get:

$$\sum_{m=1}^{M} \left(\frac{1}{|Y_L|} \sum_{\gamma \in Y_L} \delta_m(\overline{\gamma}) \right) \leq \frac{32K^3}{\epsilon}.$$

Thus, given any $\theta > 0$, (by choosing M sufficiently large) we can make sure that there exists $1 \leq m \leq M$ such that

$$\frac{1}{|Y_L|} \sum_{\gamma \in Y_L} \delta_m(\overline{\gamma}) < \theta. \tag{1}$$

Conclusion. On the scale $R \equiv r_m$, at least $1-\theta$ fraction of all vertical geodesic segments in $B(L)$ have nearly vertical images under ϕ. See Figure 3.

The difficulty is that, at this point, it may be possible that some of the (upward oriented) vertical segments in $B(L)$ may have images which are going up, and some may have images which are going down.

We think of the process we have just described as a form of "coarse differentiation". For further discussion of this process and a more general variant on the discussion in the last two subsections, see subsection 4.5.

4.3. The scheme of the proof of Theorem 2.1.

Roughly speaking, the proof proceeds in the following steps:

Step 1. *For all $\theta > 0$ there exists L_0 such that for any box $B(L)$ where $L \geq L_0$, there exists $0 \ll r \ll R \ll L_0$ such that for the tiling:*

$$B(L) = \bigsqcup_{i=1}^{N} B_i(R)$$

there exists $I \subset \{1, \dots, N\}$ *with* $|I| \geq (1-\theta)N$ *and for each* $i \in I$ *there exists a height-respecting map* $\hat{\phi}_i : B_i(R) \to \mathrm{Sol}$ *and a subset* $U_i \subset B_i(R)$ *with* $|U_i| \geq (1-\theta)|B_i(R)|$ *such that*

$$d(\phi|_{U_i}, \hat{\phi}_i) = O(r).$$

Roughly, Step 1 asserts that every sufficiently large box can be tiled into small boxes, in such a way that for most of the small boxes $B_i(R)$, the restriction of ϕ to $B_i(R)$ agrees, on most of the measure of $B_i(R)$, with a height-respecting map $\hat{\phi}_i : B_i(R) \to \mathrm{Sol}$. There is no assertion in Step 1 that the height-respecting maps $\hat{\phi}_i$ on different small boxes match up to define a height-respecting map on most of the measure on $B(L)$; the main difficulty is that some of the $\hat{\phi}_i$ may send the "up" direction to the "down" direction, while other $\hat{\phi}_i$ may preserve the up direction.

Step 1 follows from a version of (1) and some geometric arguments using Lemma 3.1. The point is that any ϵ-monotone quasi-geodesic is close to a vertical geodesic by Lemma 4.2. By the averaging argument in subsection 4.2, we find a scale R at which most segments have ϵ-monotone image under ϕ. More averaging implies that on most boxes $B_i(R)$ most geodesic segments joining the top of the box to the bottom of the box have ϵ-monotone images. We then apply Lemma 3.1 to the images of these geodesics and use this to show that the map is roughly height preserving on each $B_i(R)$. This step also uses the geometric description of $B_i(R)$ given in the last paragraph of §3, i.e. the fact that a box is coarsely a complete bipartite graph G_R on nets in the "top" and "bottom" of the box.

Step 2. *For all* $\theta > 0$ *there exists* L_0 *such that for any box* $B(L)$ *where* $L \geq L_0$, $\exists$ *subset* $U \subset B(L)$ *with* $|U| \geq (1-\theta)|B(L)|$ *and a height-respecting map* $\hat{\phi} : B(L) \to \mathrm{Sol}$ *such that*

$$d(\phi|_U, \hat{\phi}) = O(l),$$

where $l \ll L_0$.

This is the essentially the assertion that the different maps $\hat{\phi}_i$ from Step 1 are all oriented in the same way, and can thus be replaced by one standard map $\hat{\phi} : B(L) \to \mathrm{Sol}$.

Step 2 is the most technical part of the proof. The problem here derives from exponential volume growth. In Euclidean space, given a set of almost full measure U in a box, every point in the box is close to a point in U. This is not true in Sol because of exponential volume growth. Another manifestation of this difficulty is that Sol does not have a Vitali covering lemma. The proof involves using refinements of Lemma 3.1 and further averaging on the image of ϕ.

Step 3. *The map* ϕ *is* $O(L_0)$ *from a standard map* $\hat{\phi}$.

This follows from Step 2 and some geometric arguments using variants of Lemma 3.1. The large constant, $O(L_0)$, arises because we pass to very large scales to ignore the sets of small measure that arise in Steps 1 and 2.

4.4. Remarks on the proof of Theorem 1.4 and the general case. Peng's proof of Theorem 1.4 proceeds roughly using the same strategy as the proof of Theorem 1.3. The main difference is that instead of vertical geodesics one has "vertical flats" (which are the orbits of $\mathbb{R}^k$ acting on $\mathbb{R}^k \ltimes \mathbb{R}^n$). These flats are equipped with a foliation by hyperplanes which are parallel to the kernels of the weights on $\mathbb{R}^k$ defined by the map $\rho : \mathbb{R}^k \to \mathbb{R}^n$. Peng shows that the quasi-isometry roughly preserves the vertical flats, and also the restriction of the map to a flat preserves the foliation by hyperplanes. In particular a geodesic in a vertical flat which is transverse to the root hyperplanes maps roughly to another such object. This allows Peng to show that the map roughly preserves subsets of the space whose geometry is quite similar to Sol geometry. This fact then allows her to use the geometry of the graphs G_L described at the end of §3 in her arguments.

The case $G = \mathbb{R} \ltimes N$ where N is a nilpotent group and N is equal to the exponential radical of G can already present considerable extra difficulties. In this case we can split the Lie algebra $\mathfrak{n}$ of N into an expanding subspace $\mathfrak{n}^+$ and a contracting subspace $\mathfrak{n}^-$. In the case where $[\mathfrak{n}^+, \mathfrak{n}^-] = 0$, the geometry of G is quite similar to the geometry of *Sol* and the proof that the N coset foliation is preserved can be carried out in a very similar fashion. However if $[\mathfrak{n}^+, \mathfrak{n}^-] \neq 0$, the geometry is quite different and the the quadrilateral lemma (Lemma 3.1) fails to be true. This is closely related to the fact the the graphs G_L, which are the $\mathbb{R} \ltimes N$ analogues of the graph given the same name at the end of §3, are no longer complete bipartite. One can make progress in this direction by replacing Lemma 3.1 by a sort of averaged version, but this requires the detailed study of the graphs G_L, including proving a uniform spectral gap as the size of the box L tends to infinity. This is done in [FP].

Once the $\mathbb{R} \ltimes N$ case is complete, the proof in split polycyclic case $G = \mathbb{R}^k \ltimes N$ would presumably involve incorporating the ideas of [Pe1], [Pe2]. Even the split case where $G = N_1 \ltimes N_2$ with $N_2 = \exp(G)$ will be quite similar with vertical flats replaced by vertical copies of N_1. For the general polycyclic group G, the exact sequence $1 \to \exp(G) \to G \to G/\exp(G) \to 1$ may not split, and "vertical flats" are no longer defined. However, there is a geometric splitting of the exact sequence which defines a foliation of G by sets diffeomorphic to $G/\exp(G)$ where the maps sending leaves into the space $G/\exp(G) \to G$ preserves distances up to a logarithmic error and thus the methods described here are still relevant. This geometric splitting is used by de Cornulier in his work on the asymptotic geometry of solvable Lie groups [dC1, dC2].

4.5. Remarks on coarse differentiation. If a map is differentiable, then it is locally at sub-linear error from a map which takes lines to lines. This

is roughly the conclusion of the argument above for the vertical geodesics in Sol, at least on an appropriately chosen large scale and off of a set of small measure. The ideas employed here can be extended to general metric spaces, by replacing the notion of ϵ-monotone with a more general notion of ϵ-efficient which we will describe below. The ideas in our proof are not so different from the proof(s) of Rademacher's theorem that a bilipschitz map of $\mathbb{R}^n$ is differentiable almost everywhere. In fact, our method applied to quasi-isometries of $\mathbb{R}^n$ gives roughly the same information as the application of Rademacher's theorem to the induced bilipschitz map on the asymptotic cone of $\mathbb{R}^n$ (which is again $\mathbb{R}^n$). In this context the presence of sets of small measure can be eliminated by a covering lemma argument. In the context of solvable groups, passage to the asymptotic cone is complicated by the exponential volume growth. The asymptotic cone for these groups is not locally compact, which makes it difficult to find useful notions of sets of zero or small measure there.

We now formulate somewhat loosely a more general form of the "differentiation theorem" given in subsections 4.1 and 4.2. Throughout this subsection Y will be a general metric space, though it may be most useful to think of Y as a complete, geodesic metric space. First we generalize the notion of ϵ-monotone.

Definition 4.5. *A quasigeodesic segment $\alpha : [0, L] \to Y$ is ϵ-efficient on the scale r if*

$$\sum_{j=1}^{L/r} d(\alpha(jr), \alpha((j-1)r)) \leq (1+\epsilon) d(\alpha(L), \alpha(0)).$$

The fact is that a quasi-geodesic, unless it is a $(1+\epsilon)$ quasi-geodesic, fails to be ϵ-efficient at some scale some fraction of the time. The observation embedded in subsection 4.1 is that this cannot happen everywhere on all scales and in fact cannot happen too often on too many scales.

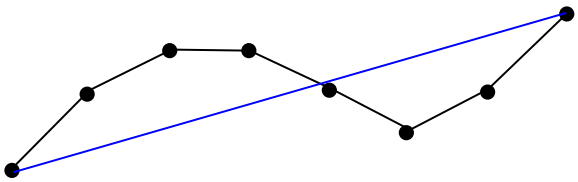


Figure 4. The definition of ϵ-efficient.

With this definition, the following variant on Lemma 4.3 becomes a tautology.

Lemma 4.6 (Subdivision II). *Given $\epsilon > 0$, there exist $r \gg C$ and $n \gg 1$ (depending on K, C and ϵ) such that any (K, C)-quasi-geodesic segment α :*

$[0,r] \to X$ which is not ϵ-efficient on scale $\frac{r}{n}$ we have:

$$\sum_{j=0}^{n-1} d\left(\alpha\left(\tfrac{(j+1)r}{n}\right), \alpha\left(\tfrac{jr}{n}\right)\right) \geq d(\alpha(0), \alpha(r)) + \frac{\epsilon r}{2K}.$$

We now state a variant of Lemma 4.4 whose proof is verbatim the proof of that lemma.

Choosing Scales: Choose $1 \ll r_0 \ll r_1 \ll \cdots \ll r_M$. In particular, $C \ll r_0$ and $r_{m+1}/r_m > n$.

Lemma 4.7. *Suppose $L \gg r_M$, and suppose $\alpha : [0, L] \to X$ is a quasi-geodesic segment. For each $m \in [1, M]$, subdivide $[0, L]$ into L/r_m segments of length r_m. Let $\delta_m(\alpha)$ denote the fraction of these segments whose images are not ϵ-efficient on scale r_{m-1}. Then,*

$$\sum_{m=1}^{M} \delta_m(\alpha) \leq \frac{4K^2}{\epsilon}.$$

Let X be a geodesic metric space. Coarse differentiation amounts to the following easy lemma.

Lemma 4.8 (Coarse Differentiation). *Let $\phi : X \to Y$ be a (K, C)-quasi-isometry. For all $\theta > 0$ there exists $L_0 \gg 1$ such that for any $L > L_0$ and any family $\mathcal{F}$ of geodesics of length L in X, there exist scales r, R with $C \ll r \ll R \ll L_0$ such that if we divide each geodesic in $\mathcal{F}$ into subsegments of length R, then at least $(1 - \theta)$ fraction of these subsegments have images which are ϵ-efficient at scale r.*

This lemma and its variants seem likely to be useful in other settings. In fact, the lemma holds only assuming that ϕ is coarsely lipschitz. A map $\phi : X \to Y$ is a (K, C) *coarsely lipschitz* if $d_Y(\phi(x_1), \phi(x_2)) \leq K d_X(x_1, x_2) + C$. We now describe the relation to taking derivatives and also to the process of taking a "derivative at infinity" of a quasi-isometry by passing to asymptotic cones.

We first discuss the case of maps $\mathbb{R}^n \to \mathbb{R}^n$. Suppose $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is a quasi-isometry. Suppose one chooses a net N on the unit circle and takes $\mathcal{F}$ to be the set of all lines of length L in a large box, whose direction vector is in N. Lemma 4.8 applied to $\mathcal{F}$ then states that most of these lines, on the appropriate scale, map under ϕ close to straight lines, which implies that the map ϕ (in a suitable box) can be approximated by an affine map. Thus, in this context, Lemma 4.8 is indeed analogous to differentiation (or producing points of differentiability).

An alternative approach for analyzing quasi-isometries $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is to pass to the asymptotic cone to obtain a bilipschitz map $\tilde{\phi} : \mathbb{R}^n \to \mathbb{R}^n$ and then apply Rademacher's theorem to $\tilde{\phi}$. If one attempts to pull the information this yields back to ϕ one gets statements that are similar to those one would obtain directly using Lemma 4.8. This is not surprising, since averaging arguments like

those used in the proof of Lemma 4.8 are implicit in the proofs of Rademacher's theorem.

Passing to the asymptotic cone has obvious advantages because it allows one to replace a (K, C) quasi-isometry from X to Y with a $(K, 0)$-quasi-isometry (i.e. a bilipschitz map) from the asymptotic cone of X to the asymptotic cone of Y. One can then try to use analytic techniques to study the bilipshitz maps. However, a major difficulty which occurs is that the asymptotic cones are typically not locally compact and notions of measure and averaging on such spaces are not clear. This difficulty arises as soon as one has exponential volume growth. In particular it is not clear if there is a useful version of Rademacher's theorem for the asymptotic cones of the spaces which we consider in this paper.

The main advantage of Lemma 4.8 compared to the asymptotic cone approach is that the averaging is done on the (typically locally compact) space X, i.e. the domain of the quasi-isometry ϕ. In other words, we construct a "coarse derivative" without first passing to a limit to get rid of the additive constant. In particular, the information we obtain about Sol and other solvable groups by coarse differentiation is not easily extracted by passage to the asymptotic cone.

We remark again that Lemma 4.8 applies to any quasi-isometric embedding (or any uniform embedding) between any two metric spaces X and Y. However its usefulness clearly depends on the situation.

The coarse differentiation approach is closely related to results proved the method of the "iterated midpoint" which is well-known in the theory of Banach spaces, see e.g. [B],[BL], [JLS], [M], [Pr], [BJLPS]. Some results of some of those papers also have a similar flavor, resulting in points where a map between Banach spaces is ϵ-Frechet differentiable, i.e. that the map is sublinear distance from an affine map at some scale. The main difference in proofs is that in our setting it is possible to average the inequality as described in §4.2 to obtain some control on a set of large (but not full) measure.

4.6. Deduction of rigidity results. In our setting, the deduction of rigidity results from the classification of quasi-isometries follows a fairly standard outline that is similar to one used for semisimple groups as well as for certain solvable groups in [FM2, FM3, MSW]. As this is standard, we will say relatively little about it. Some of these ideas go back to Mostow's original proof of Mostow rigidity [Mo1, Mo3] and have been developed further by many authors.

Given a group Γ any element of γ in Γ acts on Γ by isometries by left multiplication L_γ. If X is a metric space and $\phi : \Gamma \to X$ is a quasi-isometry, we can conjugate each L_γ to a self quasi-isometry $\phi \circ L_\gamma \circ \phi^{-1}$ of X. This induces a homomorphism of $\Phi : \Gamma \to \mathrm{QI}(X)$. Here $\mathrm{QI}(X)$ is the group of quasi-isometries of X modulo the subgroup of quasi-isometries a bounded distance from the identity. The approach we follow is to use Φ to define an action of Γ on a "boundary at infinity" of the space X. All theorems are then proven by studying

the dynamics of this "action at infinity." We are ignoring many important technical points here, such as why Φ has finite kernel and why $\mathrm{QI}(X)$ acts on either X or the boundary at infinity of X.

The deduction of Theorem 1.3 from Theorem 2.1 was known to Farb and Mosher [FM2, FM4]. The action at infinity is studied using a variant of a theorem of Hinkannen due to Farb and Mosher [H, FM2, FM4]. In the context of Theorem 1.4, we deduce the result from Theorem 2.2 using results from the dissertation of Tullia Dymarz and a further paper by Dymarz and Peng [Dy, DP]. These are variants and extensions of the results of Tukia in [Tu]. As remarked above, a proof of Conjecture 1.2 from Conjecture 2.3 will require a further generalization of these results. This generalization is already a significant and difficult problem.

5. Further Consequences

In this section we discuss some other results that are consequence either of our methods or of our results.

5.1. Geometry of Diestel–Leader graphs. In addition our methods yield quasi-isometric rigidity results for a variety of solvable groups which are not polycyclic, in particular the so-called lamplighter groups. These are the wreath products $\mathbb{Z}\wr F$ where F is a finite group. The name lamplighter comes from the description $\mathbb{Z}\wr F = F^{\mathbb{Z}} \rtimes \mathbb{Z}$ where the $\mathbb{Z}$ action is by a shift. The subgroup $F^{\mathbb{Z}}$ is thought of as the states of a line of lamps, each of which has $|F|$ states. The "lamplighter" moves along this line of lamps (the $\mathbb{Z}$ action) and can change the state of the lamp at her current position. The Cayley graphs for the generating sets $F \cup \{\pm 1\}$ depend only on $|F|$, not the structure of F. Furthermore, $\mathbb{Z}\wr F_1$ and $\mathbb{Z}\wr F_2$ are quasi-isometric whenever there is a d so that $|F_1| = d^s$ and $|F_2| = d^t$ for some s, t in $\mathbb{Z}$. The problem of classifying these groups up to quasi-isometry, and in particular, the question of whether the 2 and 3 state lamplighter groups are quasi-isometric, were well known open problems in the field, see [dlH].

Theorem 5.1. *The lamplighter groups $\mathbb{Z}\wr F$ and $\mathbb{Z}\wr F'$ are quasi-isometric if and only if there exist positive integers d, s, r such that $|F| = d^s$ and $|F'| = d^r$.*

For a rigidity theorem for lamplighter groups, see Theorem 5.2 below.

To state Theorem 5.2 as well as some other results, we need to describe a class of graphs. These are the Diestel–Leader graphs, $DL(m, n)$, which can be defined as follows: let T_1 and T_2 be regular trees of valence $m + 1$ and $n + 1$. Choose orientations on the edges of T_1 and T_2 so each vertex has n (resp. m) edges pointing away from it. This is equivalent to choosing ends on these trees. We can view these orientations at defining height functions f_1 and f_2 on the trees (the Busemann functions for the chosen ends). If one places the

point at infinity determining f_1 at the top of the page and the point at infinity determining f_2 at the bottom of the page, then the trees can be drawn as:

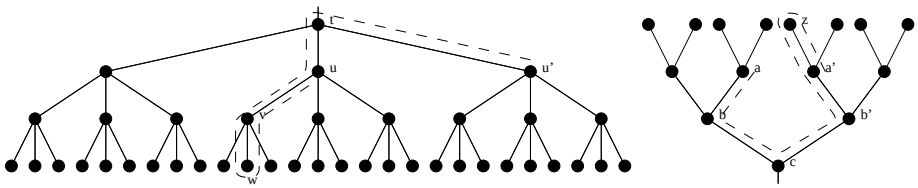


Figure 5. The trees for $DL(3,2)$. Figure borrowed from [PPS].

The graph $DL(m,n)$ is the subset of the product $T_1 \times T_2$ defined by $f_1 + f_2 = 0$. The analogy with the geometry of Sol is clear from section 3. For $n = m$ the Diestel–Leader graphs arise as Cayley graphs of lamplighter groups $\mathbb{Z} \wr F$ for $|F| = n$. This observation was apparently first made by R.Moeller and P.Neumann [MN] and is described explicitly, from two slightly different points of view, in [Wo2] and [W]. We prove the following:

Theorem 5.2. *Let Γ be a finitely generated group quasi-isometric to the lamplighter group $\mathbb{Z} \wr F$. Then there exists positive integers d, s, r such that $d^s = |F|^r$ and an isometric, proper, cocompact action of a finite index subgroup of Γ on the Diestel–Leader graph $DL(d,d)$.*

Remark: The theorem can be reinterpreted as saying that any group quasi-isometric to $DL(|F|,|F|)$ is virtually a cocompact lattice in the isometry group $\mathrm{Isom}(DL(d,d)$ of $DL(d,d)$ where d is as above.

Recently, the second author, de Cornulier and Kashyp have proven some detailed results concerning the algebraic structure of cocompact lattices in $\mathrm{Isom}(DL(d,d))$. We state here one corollary of that work.

Theorem 5.3. *Let $\Gamma < \mathrm{Isom}(DL(d,d))$ be a cocompact lattice. Then Γ admits a transitive, proper action on $DL(d^n, d^n)$ for some positive n.*

The paper [dCFK] also contains many examples of lattices in $\mathrm{Isom}(DL(d,d))$ which are not weakly commensurable to lamplighters.

In [SW, Wo1], Soardi and Woess ask whether every homogeneous graph is quasi-isometric to a finitely generated group. The graph $DL(m,n)$ is easily seen to be homogeneous (i.e. it has a transitive isometry group). For $m \neq n$ its isometry group is not unimodular, and hence has no lattices. Thus there are no obvious groups quasi-isometric to $DL(m,n)$ in this case. In fact, we have:

Theorem 5.4. *There is no finitely generated group quasi-isometric to the graph $DL(m,n)$ for $m \neq n$.*

This theorem was conjectured by Diestel and Leader in [DL], where the Diestel–Leader graphs were introduced for this purpose. Note that Theorem 5.4

can be reinterpreted as the statement that for $m \neq n$, there is no finitely generated group quasi-isometric to the isometry group of $DL(m,n)$.

Recall that $DL(m,n)$ is defined as the subset of $T_{m+1}\times T_{n+1}$ where $f_m(x)+f_n(y)=0$ where f_m and f_n are Busemann functions on T_m and T_n respectively. Here we simply set $h((x,y))=f_m(x)=-f_n(y)$ which makes sense exactly on $DL(m,n)\subset T_{m+1}\times T_{n+1}$. The reader can verify that the level sets of the height function are orbits for a subgroup of $\mathrm{Isom}(DL(m,n))$.

Theorem 5.5. *Any (K,C)-quasi-isometry φ of $DL(m,n)$ is within bounded distance from a height respecting quasi-isometry $\hat{\varphi}$. Furthermore, the bound is uniform in K and C.*

Remark: We can reformulate Theorem 5.5 in terms similar to those of Theorem 2.1. Here the group $\mathrm{Bilip}(\mathbb{R})\times\mathrm{Bilip}(\mathbb{R})$ will be replaced by $\mathrm{Bilip}(X_m)\times\mathrm{Bilip}(X_n)$ for X_m (resp. X_n) the complement of a point in the (visual) boundary of T_{m+1} (resp. T_{n+1}). These can easily be seen to be the m-adic and n-adic rationals, respectively.

Note that when $m=n$, this theorem is used to prove Theorem 5.2 and when $m\neq n$ it is used to prove Theorem 5.4. The proofs in these two cases are somewhat different, the proof in the case $m=n$ being almost identical to the proof of Theorem 2.1. In the other case, the argument is complicated by the absence of metric Fölner sets, but simplifications also occur since there is no element in the isometry group that "flips" height. There is an analogue of the above results for the case of the solvable Lie groups which appears in Theorem 5.9.

Another recent dramatic development that uses Theorem 5.5 is the following result of Dymarz [Dy2].

Theorem 5.6. *Consider the two lamplighter groups $F^k\wr\mathbb{Z}$ and $F\wr\mathbb{Z}$ where $|F|=m$ and F^k is the direct product of k copies of F. Then there does not exist a bijective quasi-isometry between $F^k\wr\mathbb{Z}$ and $F\wr\mathbb{Z}$ if k is not a product of prime factors appearing in m.*

The main point of the theorem is that these two groups are quasi-isometric but not bijectively quasi-isometric. A result of Whyte says that any pair of non-amenable groups are quasi-isometric if and only if they are bijectively quasi-isometric [Wh]. Dymarz's result proves that this is no longer true for amenable groups and answers a question that had been open for over ten years.

5.2. Low dimensional topology and geometry.

We now state a theorem that is a well-known consequence of Theorem 1.3, Thurston's Geometrization Conjecture and results in [CC, Gr1, KaL1, KaL2, PW, S1, Ri]. We state it assuming that the Geometrization Conjecture is known.

Theorem 5.7. *Let M be a compact three manifold without boundary and Γ a finitely generated group. If Γ is quasi-isometric to the universal cover of M,*

then Γ is virtually the fundamental group of M', also a compact three manifold without boundary.

For more discussion of this theorem and significant progress towards classifying three manifold groups up to quasi-isometry, see work of Behrstock and Neumann [BN1, BN2].

The existence of transitive graphs not quasi-isometric to Cayley graphs as given by Theorem 5.4 gives rise to interesting surfaces with exotic properties. The surfaces are obtained simply by replacing edges in the graphs by tubes and vertices by spheres to which one attaches the tubes. This construction is used by Bonafert, Canary, Souto and Taylor to construct uniformly quasiconformally homogeneous Riemann surfaces which are not quasiconformal deformations of regular covers of closed orbifolds [BCST].

5.3. Lie groups not quasi-isometric to discrete groups. The following is a basic question:

Question 5.8. *Given a Lie group G, is there a finitely generated group quasi-isometric to G?*

It is clear that the answer is yes whenever G has a cocompact lattice. However, many solvable locally compact groups, and in particular, many solvable Lie groups do not have any lattices. The simplest examples are groups which are not unimodular. However, it is possible for Question 5.8 to have an affirmative answer even if G is not unimodular. For instance, the non-unimodular group solvable group $\left\{ \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} \middle| a > 0, b \in \mathbb{R} \right\}$ acts simply transitively by isometries on the hyperbolic plane, and thus is quasi-isometric to the fundamental group of any closed surface of genus at least 2. Thus the answer to Question 5.8 can be subtle. Our methods give:

Theorem 5.9. *Let $G = \mathbb{R} \ltimes \mathbb{R}^2$ be a solvable Lie group where the $\mathbb{R}$ action on $\mathbb{R}^2$ is defined by $z \cdot (x, y) = (e^{az}x, e^{-bz}y)$ for $a, b > 0$, $a \neq b$. Then there is no finitely generated group Γ quasi-isometric to G.*

If $a > 0$ and $b < 0$, then G admits a left invariant metric of negative curvature. The fact that there is no finitely generated group quasi-isometric to G in this case is a result of Kleiner [K], see also [Pa2]. Kleiner's result has recently been generalized by Shanmugalingam and Xie [SX]. It is possible to generalize the theorem above further using our techniques and the results in [Dy, DP]. Nilpotent Lie groups not quasi-isometric to any finitely generated group where constructed in [ET].

5.4. Distortion of embeddings and multi-commodity flow problems. The technique of coarse differentiation has also been applied by Lee and Raghavendra to a problem arising from theoretical computer science

[LR]. Their work is motivated by the multi-commodity version of min cut- max flow. This problem is known to be related to problems concerning distortion of embeddings into L^1 spaces. They use coarse differentiation to obtain bounds on the distortion of L^1 embeddings for a certain family of graphs.

References

[A1] Auslander, Louis An exposition of the structure of solvmanifolds. I. Algebraic theory. *Bull. Amer. Math. Soc.* 79 (1973), no. 2, 227–261.

[A2] Auslander, Louis An exposition of the structure of solvmanifolds. II. G-induced flows. *Bull. Amer. Math. Soc.* 79 (1973), no. 2, 262–285.

[B] Bourgain, J. Remarks on the extension of Lipschitz maps defined on discrete sets and uniform homeomorphisms. *Geometrical aspects of functional analysis (1985/86), 157–167, Lecture Notes in Math.*, 1267, Springer, Berlin, 1987.

[BJLPS] Bates, S.; Johnson, W. B.; Lindenstrauss, J.; Preiss, D.; Schechtman, G. Affine approximation of Lipschitz functions and nonlinear quotients. *Geom. Funct. Anal.* 9 (1999), no. 6, 1092–1127.

[BLPS] Benjamini, I.; Lyons, R.; Peres, Y.; Schramm, O. Group-invariant percolation on graphs. *Geom. Funct. Anal.* 9 (1999), no. 1, 29–66.

[BN1] Behrstock, Jason A.; Neumann, Walter D. Quasi-isometric classification of graph manifold groups. *Duke Math. J.* 141 (2008), no. 2, 217–240.

[BN2] Behrstock, Jason A.; Neumann, Walter D. Quasi-isometric classification of non-geometric 3-manifold groups. Preprint.

[BL] Y. Binyamini, J. Lindenstrauss. Geometric Nonlinear Functional Analysis American Mathematical Society Colloquim publications, Vol. 48.

[BCST] Bonfert–Taylor, Petra; Canary, Richard; Souto, Juan, Taylor, Edward. Exotic quasiconformally homogeneous surfaces. Preprint.

[CC] Cannon, J. W.; Cooper, Daryl A characterization of cocompact hyperbolic and finite-volume hyperbolic groups in dimension three. Trans. Amer. Math. Soc. 330 (1992), no. 1, 419–431.

[dC1] de Cornulier, Yves. Dimension of asymptotic cones of Lie groups. *J. Topol.* 1 (2008), no. 2, 342–361.

[dC2] de Cornulier, Yves. Asymptotic cones of Lie groups and cone equivalences, preprint.

[dCFK] de Cornulier, Yves; Fisher, David; Kashyp, Neeraj. Cross-wired lamplighters, in preparation.

[dlH] de la Harpe, Pierre. *Topics in geometric group theory.* Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 2000.

[DL] Diestel, Reinhard; Leader, Imre A. conjecture concerning a limit of non-Cayley graphs. *J. Algebraic Combin.* 14 (2001), no. 1, 17–25.

[Dy] Dymarz, T. Large scale geometry of certain solvable groups. to appear *GAFA*.

[Dy2] Dymarz, T. Bilipschitz equivalence is not equivalent to quasi-isometric equivalence for finitely generated groups, to appear *Duke Math. Journal*.

[DP] Dymarz, T; Peng, I. Bilipschitz maps of boundaries of certain negatively curved homogeneous spaces. Preprint

[D] Dyubina, Anna. Instability of the virtual solvability and the property of being virtually torsion-free for quasi-isometric groups. *Internat. Math. Res. Notices* 2000, no. 21, 1097–1101.

[E] Eskin, Alex. Quasi-isometric rigidity of nonuniform lattices in higher rank symmetric spaces. *J. Amer. Math. Soc.* 11 (1998), no. 2, 321–361.

[EF] Eskin, Alex; Farb, Benson. Quasi-flats and rigidity in higher rank symmetric spaces. *J. Amer. Math. Soc.* 10 (1997), no. 3, 653–692.

[EFW0] Eskin, Alex; Fisher, David; Whyte, Kevin. Quasi-isometries and rigidity of solvable groups. *Pure Appl. Math. Q.* 3 (2007), no. 4, part 1, 927–947.

[EFW1] Eskin, Alex; Fisher, David; Whyte, Kevin. Coarse differentiation of quasi-isometries I: spaces not quasi-isometric to Cayley graphs. Preprint.

[EFW2] Eskin, Alex; Fisher, David; Whyte, Kevin. Coarse differentiation of quasi-isometries II: Rigidity for Sol and Lamplighter groups. Preprint.

[ET] Elek, Gabor; Tardos, Gabor. On roughly transitive amenable graphs and harmonic Dirichlet functions. *Proc. Amer. Math. Soc.* 128 (2000), no. 8, 2479–2485.

[FS] Farb, Benson; Schwartz, Richard. The large-scale geometry of Hilbert modular groups. *J. Differential Geom.* 44 (1996), no. 3, 435–478.

[F] Farb, Benson. The quasi-isometry classification of lattices in semisimple Lie groups. *Math. Res. Lett.* 4 (1997), no. 5, 705–717.

[FM1] Farb, Benson; Mosher, Lee. A rigidity theorem for the solvable Baumslag–Solitar groups. With an appendix by Daryl Cooper. *Invent. Math.* 131 (1998), no. 2, 419–451.

[FM2] Farb, Benson; Mosher, Lee. Quasi-isometric rigidity for the solvable Baumslag–Solitar groups. II. *Invent. Math.* 137 (1999), no. 3, 613–649.

[FM3] Farb, Benson; Mosher, Lee. On the asymptotic geometry of abelian-by-cyclic groups. *Acta Math.* 184 (2000), no. 2, 145–202.

[FM4] Farb, Benson; Mosher, Lee. Problems on the geometry of finitely generated solvable groups. *Crystallographic groups and their generalizations (Kortrijk, 1999)* , 121–134, Contemp. Math., 262, Amer. Math. Soc., Providence, RI, 2000.

[FP] Fisher, D; Peng, I; Geometry of rank one solvable Lie groups, in preparation.

[Gr1] Gromov, Mikhael. Groups of polynomial growth and expanding maps. *Inst. Hautes Études Sci. Publ. Math.* No. 53 (1981), 53–73.

[Gr2] Gromov, Mikhael. Infinite groups as geometric objects. Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983), 385–392, PWN, Warsaw, 1984.

[Gr3] Gromov, M. Asymptotic invariants of infinite groups. *Geometric group theory, Vol. 2 (Sussex, 1991)*, 1–295, London Math. Soc. Lecture Note Ser., 182, Cambridge Univ. Press, Cambridge, 1993.

[Gu] Guivarc'h, Y. Théorèmes quotients pour les marches aléatoires. (French) Conference on Random Walks (Kleebach, 1979) (French), pp. 15–28, 3, *Astérisque*, 74, Soc. Math. France, Paris, 1980.

[H] Hinkkanen, A. Uniformly quasisymmetric groups. *Proc. London Math. Soc.* (3) 51 (1985), no. 2, 318–338.

[JLS] Johnson, W. B.; Lindenstrauss, J.; Schechtman, G. Banach spaces determined by their uniform structures. *Geom. Funct. Anal.* 6 (1996), no. 3, 430–470.

[KaL1] Kapovich, Michael; Leeb, Bernhard Quasi-isometries preserve the geometric decomposition of Haken manifolds. *Invent. Math.* 128 (1997), no. 2, 393–416.

[KaL2] Kapovich, M.; Leeb, B. 3-manifold groups and nonpositive curvature. *Geom. Funct. Anal.* 8 (1998), no. 5, 841–852.

[KL] Kleiner, Bruce; Leeb, Bernhard. Rigidity of quasi-isometries for symmetric spaces and Euclidean buildings. *Inst. Hautes Études Sci. Publ. Math.* No. 86, (1997), 115–197 (1998).

[K] Kleiner, Bruce. Personal communication.

[LR] Lee, James R; Raghavendra, P. Coarse differentiation and planar multiflows. To appear *Discrete and Computational Geometry*.

[M] J. Matousek. Embedding Trees into Uniformly Convex Banach Spaces. *Israel J of Math*, 1999.

[MN] Letter from R.Moeller to W.Woess, 2001.

[MSW] Mosher, Lee; Sageev, Michah; Whyte, Kevin. Quasi-actions on trees. I. Bounded valence. *Ann. of Math.* (2) 158 (2003), no. 1, 115–164.

[Mo1] Mostow, G. D. Quasi-conformal mappings in n-space and the rigidity of hyperbolic space forms. *Inst. Hautes Études Sci. Publ. Math.* No. 34 1968 53–104.

[Mo2] Mostow, G. D. Representative functions on discrete groups and solvable arithmetic subgroups. *Amer. J. Math.* 92 1970 1–32.

[Mo3] Mostow, G. D. Strong rigidity of locally symmetric spaces. Annals of Mathematics Studies, No. 78. Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo, 1973.

[Os] Osin, D. V. Exponential radicals of solvable Lie groups. *J. Algebra* 248 (2002), no. 2, 790–805.

[Pr] Preiss, D. Differentiability of Lipschitz functions on Banach spaces. *J. Funct. Anal.* 91 (1990), no. 2, 312–345.

[Pa1] Pansu, Pierre. Metriques de Carnot–Carathéodory et quasiisométries des espaces symétriques de rang un. (French) [Carnot–Caratheodory metrics and quasi-isometries of rank-one symmetric spaces] *Ann. of Math.* (2) 129 (1989), no. 1, 1–60.

[Pa2] Pansu, Pierre. Dimension conforme et sphère l'infini des variétés à courbure négative. (French) [Conformal dimension and sphere at infinity of manifolds of negative curvature] *Ann. Acad. Sci. Fenn. Ser. A I Math.* 14 (1989), no. 2, 177–212.

[Pe1] Peng, I. Coarse differentiation and quasi-isometries of a class of solvable Lie groups I. Preprint.

[Pe2] Peng, I. Coarse differentiation and quasi-isometries of a class of solvable Lie groups II. Preprint.

[PW] Papasoglu, Panos; Whyte, Kevin Quasi-isometries between groups with infinitely many ends. *Comment. Math. Helv.* 77 (2002), no. 1, 133–144.

[PPS] Yuval Peres, Gabor Pete, Ariel Scolnicov. Critical percolation on certain non-unimodular graphs, *New York J. Math.* 12 (2006), 1–18 (electronic).

[Ri] E. Rieffel, Groups coarse quasi-isometric to $\mathbb{H}^2 \times \mathbb{R}$, PhD thesis, UCLA 1993.

[Sa] Sauer, Roman. Homological Invariants and Quasi–Isometry. *Geom. Funct. Anal.* 16 (2006), no. 2, 476–515.

[S1] Schwartz, Richard Evan. The quasi-isometry classification of rank one lattices. *Inst. Hautes Études Sci. Publ. Math.* No. 82 (1995), 133–168 (1996).

[S2] Schwartz, Richard Evan. Quasi-isometric rigidity and Diophantine approximation. *Acta Math.* 177 (1996), no. 1, 75–112.

[Sh] Shalom, Yehuda. Harmonic analysis, cohomology, and the large-scale geometry of amenable groups. *Acta Math.* 192 (2004), no. 2, 119–185.

[SX] Shanmugalingam,Nageswari; Xie,Xiangdong. A Rigidity Property of Some Negatively Curved Solvable Lie Groups. Preprint.

[SW] Soardi, Paolo M.; Woess, Wolfgang. Amenability, unimodularity, and the spectral radius of random walks on infinite graphs. *Math. Z.* 205 (1990), no. 3, 471–486.

[Tu] Tukia, Pekka. On quasiconformal groups. *J. Analyse Math.* 46 (1986), 318–346.

[Wh] Whyte, Kevin. Amenability, bi-Lipschitz equivalence, and the von Neumann conjecture. *Duke Math. J.* 99 (1999), no. 1, 93–112.

[Wo1] Woess, Wolfgang. Topological groups and infinite graphs. *Directions in infinite graph theory and combinatorics (Cambridge, 1989). Discrete Math.* 95 (1991), no. 1–3, 373–384.

[Wo2] Woess, Wolfgang. Lamplighters, Diestel–Leader graphs, random walks, and harmonic functions, *Combinatorics, Probability & Computing* 14 (2005) 415–433.

[W] Wortman, Kevin. A finitely presented solvable group with small quasi-isometry group. *Michigan Math. J.* 55 (2007), no. 1, 3–24.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Rational Cherednik Algebras

Iain G. Gordon*

Abstract

We survey a number of results about the rational Cherednik algebra's representation theory and its connection to symplectic singularities and their resolutions.

Mathematics Subject Classification (2010). Primary 16G, 17B; Secondary 20C, 53D.

Keywords. Cherednik algebra, symplectic singularity, hamiltonian reduction.

1. Introduction

This paper explores some rational Cherednik algebra representation theory and its interaction with constructions in algebraic geometry with a symplectic flavour. Although the rational Cherednik algebras were constructed as degenerations of Cherednik's double affine Hecke algebra and so have many links with the theory as developed there, see [18], it turns out that a connection with the theory of symplectic resolutions, and particularly Hilbert schemes, has played a particularly important role too. Such a connection was already foreseen at the birth of the algebras, and over the last decade the subject has developed significantly in this direction. There have been constructions of symplectic resolutions via moduli spaces of representations and also localisation theorems from the categories of representations to sheaves on quantisations of the resolutions. Since symplectic resolutions turn up remarkably often in representation theory this in turn has led to the study of the geometry and algebra of such resolutions in general. Here the Cherednik algebras are key examples helping to form the subject. The goal of this brief survey is to present a little of this.

We completely omit lots of interesting aspects of Cherednik algebras, including realisations as Hecke algebras for double loop groups, as equivariant

*The author is grateful for the full financial support of EPSRC grant EP/007632.

School of Mathematics and Maxwell Institute of Mathematics, University of Edinburgh, JCMB, King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, Scotland, UK.
E-mail: igordon@ed.ac.uk.

K-groups of affine flag manifolds, as Hall algebras of elliptic curves and via the equivariant K-theory of the Hilbert scheme. There are, however, a number of surveys on rational Cherednik algebras where many more details can be found, [30], [25], [62], [42], [71], [28].

The structure of the article is as follows. We begin in Section 2 by defining rational Cherednik algebras. In the third section we discuss symplectic singularities, representation theory at $t = 0$, and the existence of symplectic resolutions of orbit singularities. In Section 4 we explain the KZ functor, induction and restriction functors, and results on supports of representations. In the final section we present a number of different approaches to localisation of the rational Cherednik algebras of type A to the Hilbert scheme of points on the plane.

2. Definitions

Rational Cherednik algebras are defined for any finite complex reflection group W.

Definition 1. A *complex reflection group* W is a group acting on a finite dimensional complex vector space $\mathfrak{h}$ that is generated by complex reflections: non-trivial elements that fix a complex hyperplane in $\mathfrak{h}$ pointwise. We say W is *irreducible* if $\mathfrak{h}$ is an irreducible representation of W.

Such groups, which include the finite Coxeter groups, play a major role in Lie theory and invariant theory, as well as appearing in many other fields. The irreducible complex reflection groups were classified in [65]: one infinite family appears, labelled $G(d, e, n)$ where d, e, n are positive integers such that e divides d (the Coxeter groups of type A_{n-1}, B_n and D_n are $G(1, 1, n), G(2, 1, n)$ and $G(2, 2, n)$ respectively); there are 34 exceptional cases.

Given a complex reflection group W, let $\mathcal{S}$ denote its set of complex reflections, and for $s \in \mathcal{S}$ let $\alpha_s \in \mathfrak{h}^*$ have kernel the hyperplane fixed by s. We set

$$\mathbf{k} = \mathbb{C}[\mathbf{t}, \mathbf{c}_s : s \in \mathcal{S},\ \mathbf{c}_s = \mathbf{c}_{s'} \text{ if } s \text{ and } s' \text{ are conjugate in } W].$$

Definition 2 (Etingof-Ginzburg, [27]). The *rational Cherednik algebra* $\mathsf{H}_{\mathbf{k}}(W)$ is the $\mathbf{k}$-subalgebra of $\mathrm{End}_{\mathbf{k}}(\mathbf{k}[\mathfrak{h}])$ generated by the following operators:

- the action of $w \in W$
- multiplication by each $p \in \mathfrak{h}^* \subset \mathbf{k}[\mathfrak{h}]$
- for each $y \in \mathfrak{h}$, $T_y := \mathbf{t}\partial_y + \sum_{s \in \mathcal{S}} \mathbf{c}_s \alpha_s(y)\alpha_s^{-1}(s - 1)$, where ∂_y is the $\mathbf{k}$-linear derivative on $\mathbf{k}[\mathfrak{h}]$ in the direction of y.

The operators T_y are called Dunkl operators (these were introduced by Dunkl for Coxeter groups [22]; for complex reflection groups see [24]). Remarkably, the Dunkl operators commute with one another – the subalgebra of $\mathsf{H}_{\mathbf{k}}(W)$ they generate is isomorphic to $\mathbf{k}[\mathfrak{h}^*]$. This is part of the following "PBW theorem".

Theorem 2.1 ([27]). *There is a* $\mathbf{k}$*-module isomorphism*

$$\mathsf{H}_{\mathbf{k}}(W) \xrightarrow{\sim} \mathbf{k}[\mathfrak{h}] \otimes_{\mathbf{k}} \mathbf{k}[W] \otimes_{\mathbf{k}} \mathbf{k}[\mathfrak{h}^*]$$

where each tensorand is a subalgebra of $\mathsf{H}_{\mathbf{k}}(W)$*.*

Specialisation $\mathbf{k} \longrightarrow \mathbb{C}$ to parameters $t \in \mathbb{C}$ and $c \in \mathbb{C}[\mathcal{S}]^{\mathrm{ad}W}$ leads to the rational Cherednik algebra $\mathsf{H}_{t,c}(W)$, a $\mathbb{C}$-algebra. The PBW theorem says that the $\mathsf{H}_{t,c}(W)$ are deformations of $\mathsf{H}_{0,0}(W) \cong \mathbb{C}[\mathfrak{h} \times \mathfrak{h}^*] \rtimes W$, the coordinate ring of the quotient stack $[(\mathfrak{h} \times \mathfrak{h}^*)/W]$.

Definition 3. Let $e = |W|^{-1} \sum_{w \in W} w \in \mathbb{C}W$, the trivial idempotent. The *spherical Cherednik algebra* $\mathsf{U}_{\mathbf{k}}(W)$ is the $\mathbf{k}$-algebra $e\mathsf{H}_{\mathbf{k}}(W)e$.

Specialisation this time leads to the family of $\mathbb{C}$-algebras $\mathsf{U}_{t,c}(W)$. These are deformations of $\mathsf{U}_{0,0}(W) = e(\mathbb{C}[\mathfrak{h} \times \mathfrak{h}^*] \rtimes W)e \cong \mathbb{C}[\mathfrak{h} \times \mathfrak{h}^*]^W$, the coordinate ring of the orbit space $(\mathfrak{h} \times \mathfrak{h}^*)/W$.

If $\lambda \in \mathbb{C}^*$ then $\mathsf{H}_{t,c}(W) \cong \mathsf{H}_{\lambda t,\lambda c}(W)$ and $\mathsf{U}_{t,c}(W) \cong \mathsf{U}_{\lambda t,\lambda c}(W)$ so we can assume that either $t = 0$ or $t = 1$. There is now a dichotomy: $\mathsf{U}_{0,c}(W)$ is commutative, but $\mathsf{U}_{1,c}(W)$ has a trivial centre; similarly, $\mathsf{H}_{0,c}(W)$ is a finite module over its centre, but the centre of $\mathsf{H}_{1,c}(W)$ is trivial. See [27] and [16].

Remarks 1. If $W = \mathbb{Z}_2$, the cyclic group of order 2, then $U_{1,c}(W) \cong U(\mathfrak{sl}_2)/(\Omega - \lambda(c))$ where Ω is the Casimir and $\lambda(c)$ a weight depending quadratically on c. More generally, for $W = \mathbb{Z}_d = G(d, 1, 1)$ the spherical algebras were studied in the context of generalisations of the above Lie theoretic quotient and also as (commutative and noncommutative) deformations of the kleinian singularity of type A_{d-1}. For these W the algebras $H_{t,c}(W)$ were then introduced by Crawley-Boevey and Holland in [19] where they also studied the other kleinian singularities.

3. Resolutions and Deformations

The varieties $(\mathfrak{h} \times \mathfrak{h}^*)/W$ appearing above have symplectic singularities, a class of examples with rich algebraic, geometric and representation theoretic properties.

Definition 4. (Beauville, [1]) Let X be a normal affine variety over $\mathbb{C}$ that admits a symplectic 2-form ω on its smooth locus $\mathsf{sm}(X)$. We say that X has *symplectic singularities* if for any resolution of singularities $\pi : \widetilde{X} \to X$ the 2-form induced on $\pi^{-1}(\mathsf{sm}(X))$ extends to a regular 2-form on $\widetilde{X}$. If, in addition, there is a contracting $\mathbb{C}^*$-action on X with unique fixed point and such that $\lambda \cdot \omega = \lambda^n \omega$ for some positive integer n and for all $\lambda \in \mathbb{C}^*$, then we say that X has *contracting symplectic singularities*.

The paper [1] shows that $(\mathfrak{h} \times \mathfrak{h}^*)/W$ has contracting symplectic singularities: its smooth locus is the set of orbits of cardinality $|W|$ and the symplectic form

on them is inherited from the natural W-equivariant symplectic form on $\mathfrak{h} \times \mathfrak{h}^*$; dilation on the vector space $\mathfrak{h} \times \mathfrak{h}^*$ produces the $\mathbb{C}^*$-action. There are many other examples of contracting symplectic singularities in representation theory: $\mathcal{N}(\mathfrak{g})$, the nullcone of reductive Lie algebra $\mathfrak{g}$; the normalisation of the closure of a nilpotent orbit in $\mathcal{N}(\mathfrak{g})$; Slodowy's transverse slices to nilpotent orbits in $\mathcal{N}(\mathfrak{g})$; hypertoric varieties; affine Nakajima quiver varieties.

A systematic study of symplectic singularities in [48] shows they have a canonical stratification by *finitely* many symplectic leaves.

Definition 5. Suppose X has symplectic singularities. A resolution $\pi : \widetilde{X} \longrightarrow X$ is called a *symplectic resolution* if the extension of the 2-form to $\widetilde{X}$ is non-degenerate.

We have that $\pi : \widetilde{X} \longrightarrow X$ is a symplectic resolution if and only if it is a crepant resolution, see [33]. Thus, since the canonical bundle of $\widetilde{X}$ is obviously trivial in this case, the bounded derived category of coherent sheaves on $\widetilde{X}$ is of significant interest in algebraic geometry, see [49] for important results in this direction. Moreover, the Springer resolution $\pi : T^*(G/B) \longrightarrow \mathcal{N}(\mathfrak{g})$, resolutions of kleinian singularities, and many Nakajima quiver varieties are symplectic resolutions, so the notion pervades geometric representation theory.

If X has symplectic singularities then ω defines a Poisson bracket on $\mathcal{O}_X$. A Poisson deformation of X is simultaneously a deformation of the variety X *and* its bracket. There is a satisfying theory of such Poisson deformations: building on work of Ginzburg-Kaledin, [38], and using the minimal model programme, Namikawa proved

Theorem 3.1 ([59]). *Let X have contracting symplectic singularities. The following are equivalent:*

1. *X has a smooth Poisson deformation,*
2. *X has a symplectic projective resolution.*

The Grothendieck-Springer resolution illustrates this theorem:

$$\begin{array}{ccccc} T^*(G/B) & \xrightarrow{\pi} & \mathcal{N}(\mathfrak{g}) & \longrightarrow & 0 \\ \downarrow & & \downarrow & & \downarrow \\ G \times_B \mathfrak{b} & \longrightarrow & \mathfrak{g} & \xrightarrow{\delta} & \mathfrak{g}//G \end{array}$$

Here $T^*(G/B)$ is a symplectic resolution of $\mathcal{N}(\mathfrak{g})$, whilst the generic fibre of δ is G/T, a Poisson smoothing of $\mathcal{N}(\mathfrak{g})$. This also illustrates that the resolution deforms as well, a general fact for symplectic resolutions of contracting symplectic singularities.

The Grothendieck-Springer resolution is the source of a lot of remarkable representation theory; it is hoped that there is an equally rich picture around

other symplectic singularities. Rational Cherednik algebras have proved very useful in understanding this: they are related to $(\mathfrak{h}\times\mathfrak{h}^*)/W$ in the way that the enveloping algebra of $\mathfrak{g}$ is related to $\mathcal{N}(\mathfrak{g})$, but there are several new phenomena which lead to many interesting and sometimes surprising developments.

Recall that the spherical algebra $\mathsf{U}_{0,c}(W)$ is commutative for all choices of c. In fact $\mathsf{U}_{0,c}(W)\cong Z(\mathsf{H}_{0,c}(W))$, the centre of $\mathsf{H}_{0,c}(W)$, [27]. Let $\mathsf{X}_c(W)=\mathrm{Spec}(\mathsf{U}_{0,c})$. These varieties are Poisson deformations of $\mathsf{X}_0(W)=(\mathfrak{h}\times\mathfrak{h}^*)/W$, the Poisson structure on $\mathsf{U}_{0,c}(W)$ being inherited from the commutator on the flat family $\mathbb{C}[\mathbf{t}]\longrightarrow\mathsf{U}_{\mathbf{t},c}(W)$: $\{F|_{\mathbf{t}=0},G|_{\mathbf{t}=0}\}=(t^{-1}[F,G])|_{\mathbf{t}=0}$ for $F,G\in\mathsf{U}_{\mathbf{t},c}(W)$. Thus the rational Cherednik algebras provide a family of Poisson deformations over $\mathbb{C}[\mathcal{S}]^{\mathrm{ad}W}$ as well as a coherent sheaf $\mathcal{R}_c(W)$ on $\mathsf{X}_c(W)$, corresponding to the $\mathsf{U}_{0,c}(W)$-module $e\mathsf{H}_{0,c}(W)$, whose endomorphism ring is $\mathsf{H}_{0,c}(W)$.

If L is an irreducible representation of $\mathsf{H}_{0,c}(W)$, then $Z(\mathsf{H}_{0,c}(W))$ acts by scalars on it, and we have a surjective map

$$\chi_c:\mathsf{Irrep}(\mathsf{H}_{0,c}(W))\longrightarrow\mathsf{X}_c(W).$$

This is finite-to-one and from general principles of noncommutative algebra, we can use χ to study the singularities of $\mathsf{X}_c(W)$.

The prototype of such a principle is the theorem that the "Azumaya locus equals the smooth locus". Since $\mathsf{H}_{0,c}(W)$ is a finite module over its centre, there is an upper bound on the complex dimension of an irreducible $\mathsf{H}_{0,c}(W)$-representation; the Azumaya locus is by definition the set of maximal dimensional irreducible representations. It transpires that χ is one-to-one precisely on this locus, and that its image is the smooth locus of $\mathsf{X}_c(W)$, [27]. Over this locus, $\mathcal{R}_c(W)$ is actually a vector bundle of rank $|W|$, the maximal dimension of an irreducible, and we then deduce that over this locus $\mathsf{H}_{0,c}(W)$ is a matrix ring over $\mathcal{O}_{\mathsf{sm}(\mathsf{X}_c(W))}$.

Each $\mathsf{X}_c(W)$ has symplectic singularities and so is stratified by finitely many symplectic leaves. Thanks to [16] the irreducible representation theory of $\mathsf{H}_{0,c}(W)$ is constant along each leaf; elegant work of Losev, [53], and of Bellamy, [5], reduces the problem of studying a general leaf to a leaf of dimension 0, i.e. a point.

Remarks 2. There are general theorems on algebras that are finite modules over their centres that imply the Azumaya result mentioned here, [51], [15], [67]. Common to all these results is that the Azumaya locus should be relatively large (e.g. of codimension two) in the spectrum of the centre. Symplectic-like structures usually ensure this, since symplectic leaves are always even dimensional. One sees this in many Lie theoretic examples: the result holds for enveloping algebras of reductive Lie algebras in positive characteristic because of the symplectic structure on coadjoint orbits; it fails for affine Hecke algebras because there is no non-degenerate enough Poisson structure on their centre.

Similarly, passing from an arbitrary leaf to a point by considering transverse slices is a normal tactic. For instance, Premet's work on Lie algebras in positive characteristic, [61], shows that along each coadjoint orbit the representation theory is equivalent to that of the associated finite W-algebra, which is attached to the transverse slice of the orbit, and in which the orbit shrinks to a point.

There is an embedding of $R := \mathbb{C}[\mathfrak{h}]^W \otimes \mathbb{C}[\mathfrak{h}^*]^W$ into $\mathsf{U}_{0,c}(W)$, and hence a (finite) morphism $\Upsilon_c : \mathsf{X}_c(W) \to \mathfrak{h}/W \times \mathfrak{h}^*/W$. If a point $x \in X_c$ is a symplectic leaf, then it must belong to the fibre $\Upsilon_c^{-1}(0)$. By studying this fibre and applying Theorem 3.1 one can prove the following.

Theorem 3.2 ([39], [38], [3]). *For some (and hence for generic) $c \in \mathbb{C}[\mathcal{S}]^{adW}$ the variety $\mathsf{X}_c(W)$ is smooth if and only if $W = G(d,1,n)$ or $W = G_4$. It follows that $(\mathfrak{h} \times \mathfrak{h}^*)/W$ admits a symplectic projective resolution if and only if W is one of these groups.*

For $W = G(d,1,n)$ we obtain a symplectic resolution as follows, [72]. Let $Y = \mathbb{C}^2/\mathbb{Z}_d$ be the kleinian singularity of type A_{d-1} and let $\widetilde{Y}$ be its minimal resolution. Then

$$\pi : \mathsf{Hilb}^n(\widetilde{Y}) \to \mathsf{Sym}^n(\widetilde{Y}) \to \mathsf{Sym}^n(Y) = (\mathfrak{h} \times \mathfrak{h}^*)/W \tag{1}$$

is a symplectic projective resolution. This is a quiver variety; variation of GIT gives several other resolutions.

The group $W = G_4$ is an exceptional complex reflection group in the list of [65]. Two symplectic resolutions of $(\mathfrak{h} \times \mathfrak{h}^*)/W$, a four dimensional variety, are given in [52]. It remains to see whether these can be adequately described by some quiver variety construction.

The reduction of Losev and Bellamy shows that it is crucial to understand $\Upsilon_c^{-1}(0)$ and the corresponding representations of $\mathsf{H}_{0,c}(W)$. The points in $\Upsilon_c^{-1}(0)$ are equivalent to blocks in the *restricted rational Cherednik algebra* $\mathsf{H}_{0,c}(W) \otimes_R \mathbb{C}$. The irreducible representations of this algebra are labelled by the irreducible representations of W. It follows that the fibres of χ_c above $\Upsilon_c^{-1}(0)$ induce a partition of $\mathsf{Irrep}(W)$ which depends crucially on the parameter $c \in \mathbb{C}[\mathcal{S}]^{\mathrm{ad}W}$. It is conjectured, [44] and [54], that this partition essentially agrees with the decomposition of the cyclotomic Hecke algebra of W (specialised according to the choice of c) into blocks – these are called Rouquier families. Furthermore the dimension of the scheme theoretic fibre of $\Upsilon_c^{-1}(0)$ at this point should be the dimension of the corresponding Hecke algebra block. The first claim of this conjecture is confirmed for $W = G(d,e,n)$, [44] and [4], and the second claim holds whenever the given point of $\Upsilon_c^{-1}(0)$ is smooth in $\mathsf{X}_c(W)$. There is, however, no conceptual understanding of why this should be so; in particular in the Weyl group case, this suggests a link between the singularities of the spaces $\mathsf{X}_c(W)$ and Kazhdan-Lusztig theory.

4. Representations and Hecke Algebras

The algebra $\mathsf{H}_{1,c}(W)$ is sensitive to the choice of parameter $c \in \mathbb{C}[\mathcal{S}]^{\mathrm{ad}W}$: for most choices $\mathsf{H}_{1,c}(W)$ is simple; for infinitely many values of c, however, there are finite dimensional representations, and hence two-sided ideals of finite codimension. Thus we need a robust category of representations to study. Motivated by Theorem 2.1 we have the following definition, [24].

Definition 6. $\mathcal{O}_c(W)$ is the full subcategory of finitely generated $\mathsf{H}_{1,c}(W)$-modules that are locally nilpotent for the action of $\mathfrak{h} \subset \mathbb{C}[\mathfrak{h}^*] \subset \mathsf{H}_{1,c}(W)$.

This an analogue of the BGG category $\mathcal{O}$ for semisimple Lie algebras. There are related versions of $\mathcal{O}_c(W)$ where $\mathfrak{h}$ acts by non-zero eigenvalues, but [10] shows that such categories are equivalent to $\mathcal{O}_c(W')$ for some subgroup W' of W.

There is an isomorphism $\mathsf{H}_{1,0}(W) \cong D(\mathfrak{h}) \rtimes W$, the ring of W-equivariant differential operators on $\mathfrak{h}$. Hence $\mathcal{O}_0(W)$ corresponds to W-equivariant holonomic $\mathcal{D}(\mathfrak{h})$-modules whose support equals $\mathfrak{h}$, in other words to finite rank W-equivariant vector bundles on $\mathfrak{h}$ with trivial connection. This category is equivalent to the category of finite dimensional $\mathbb{C}[W]$-modules: $V \in \mathbb{C}[W]\text{-mod} \mapsto \Delta_0(V) := H_{1,0}(W) \otimes_{\mathbb{C}[\mathfrak{h}]\rtimes W} V \cong \mathbb{C}[\mathfrak{h}] \otimes V$.

In general, we can define standard modules $\Delta_c(V) \in \mathcal{O}_c(W)$, but they may no longer be the only objects in the category. If $V \in \mathsf{Irrep}(W)$ then $\Delta_c(V)$ does, however, have a unique irreducible quotient, $L_c(V)$, and $\mathcal{O}_c(W)$ becomes a highest weight category with these standard and irreducible objects. It is an important open problem to determine the composition multiplicities $[\Delta_c(V) : L_c(V')]$ for $V, V' \in \mathsf{Irrep}(W)$.

Definition 2 shows that $\mathsf{H}_{1,c}(W)[\alpha_s^{-1} : s \in \mathcal{S}] \cong D(\mathfrak{h}_{\mathrm{reg}}) \rtimes W$ where $\mathfrak{h}_{\mathrm{reg}} = \{z \in \mathfrak{h} : \alpha_s(z) \neq 0 \text{ for all } s \in \mathcal{S}\}$, the subset of $\mathfrak{h}$ on which W acts freely. Hence, on restricting to $\mathfrak{h}_{\mathrm{reg}}$, we may pass from $\mathcal{O}_c(W)$ to a category of W-equivariant bundles on $\mathfrak{h}_{\mathrm{reg}}$ with flat connections, which in turn corresponds to some category of representations of the fundamental group $\pi_1(\mathfrak{h}_{\mathrm{reg}}/W)$, a generalised Artin braid group. These representations satisfy certain Hecke-type relations.

Theorem 4.1 ([37]). *There is an exact and essentially surjective functor*

$$\mathsf{KZ}_c : \mathcal{O}_c(W) \longrightarrow \mathcal{H}_q(W)\text{-mod}$$

where $\mathcal{H}_q(W)$ denotes the (topological) Hecke algebra of W at parameter $q = \exp(2\pi i c)$ (see [14] for a definition).

This functor has many good properties. In particular it generally restricts to an equivalence on $\mathcal{O}_c(W)^{\Delta}$, the subcategory of objects that have a filtration by standard objects. Remarkably, in [63], Rouquier shows that the data of such a functor on a highest weight category together with a compatible partial order on its simple objects determines the highest weight category up to equivalence.

For $W = S_n$, there is a Schur functor $S_q(n)$-mod $\longrightarrow \mathcal{H}_q(S_n)$-mod from the q-Schur algebra, $S_q(n)$, which has analogous properties to KZ_c, see for instance [21]. Thus Rouquier's result above implies that there is an equivalence of categories between $\mathcal{O}_c(S_n)$ and $S_q(n)$-mod which sends standard modules to Weyl modules (or dual Weyl modules if c is a negative number). In particular, since the decomposition numbers are known for the q-Schur algebra, [69], we can describe the composition multiplicities $[\Delta_c(V) : L_c(V')]$ in this case in terms of parabolic Kazhdan-Lusztig polynomials of type $\widehat{A}$.

For $W = G(d, 1, n)$ and for $c \in \mathbb{C}[\mathcal{S}]^{\mathrm{ad}W}$ in a certain cone, one can show similarly that $\mathcal{O}_c(W)$ is Morita equivalent to a cyclotomic q-Schur algebra. A conjecture of Yvonne, [73], describes $[\Delta_c(V) : L_c(V')]$ in terms of a canonical basis of a level d Fock space, introduced in [68]. This conjecture is generalised to more general $c \in \mathbb{C}[\mathcal{S}]^{\mathrm{ad}W}$ in [63].

Remarks 3. There is another approach to the decomposition numbers of $\mathcal{O}_c(S_n)$ by Suzuki, [66]. Using conformal coinvariants, he constructs a functor from the Kazhdan-Lusztig category of modules for the affine Lie algebra of type $\widehat{A}$ at negative level to $\mathcal{O}_c(S_n)$. This produces an appropriate equivalence which again yields the above decomposition numbers. This functor is generalised to the $G(d, 1, n)$ case in [70] using conformal coinvariants twisted by a cyclic group action, but the corresponding decomposition numbers do not yet follow.

KZ_c is not generally a category equivalence since the passage from $\mathfrak{h}$ to $\mathfrak{h}_{\mathrm{reg}}$ kills any object of $\mathcal{O}_c(W)$ supported on $\mathfrak{h} \setminus \mathfrak{h}_{\mathrm{reg}}$, the union of reflecting hyperplanes of reflections in W. The support of an irreducible object is always a W-orbit of an intersection of reflecting hyperplanes, [35], so has, up to conjugacy, a parabolic subgroup W' attached to it by taking the stabiliser of a generic point in the intersection of these hyperplanes. Despite there usually being no non-trivial homomorphism from $\mathsf{H}_{1,c}(W')$ to $\mathsf{H}_{1,c}(W)$, Bezrukavnikov-Etingof have proved the following theorem by completing the rational Cherednik algebras at a point in the intersection of the relevant hyperplanes.

Theorem 4.2 ([10]). *Let $x \in \mathfrak{h}$ with stabiliser W_x. There are induction and restriction functors*

$$\mathcal{O}_c(W) \underset{\mathsf{Ind}_x}{\overset{\mathsf{Res}_x}{\rightleftarrows}} \mathcal{O}_c(W_x)$$

Up to isomorphism, these functors are independent of the choice of $x \in \mathfrak{h}_{reg}^{W_x} := \{z \in \mathfrak{h} : W_z = W_x\}$.

The isomorphism of functors is not canonical, and so the functor Res_x has monodromy on $\mathfrak{h}_{\mathrm{reg}}^{W_x}$. If $x \in \mathfrak{h}_{\mathrm{reg}}$ so that $W_x = 1$, the monodromy of the functor $\mathsf{Res}_x : \mathcal{O}_c(W) \longrightarrow \mathcal{O}_c(W_x) = \mathbb{C}$-mod recovers KZ_c. These functors are crucial to understanding $\mathcal{O}_c(W)$ and restriction to non-generic points preserves infor-

mation killed by KZ_c. In [64], Shan has refined these functors to produce a crystal structure on the irreducible objects in $\mathcal{O}_c(G(d,1,n))$-modules (where n varies); this crystal is isomorphic to the one attached to the canonical basis of the level d Fock space above.

In studying induction and restriction it is important to know the support of representations. Etingof uses the Macdonald-Mehta integral for Weyl groups in [26] to give a beautiful description of the support of $L_c(\mathsf{triv})$, generalising the work of [70] which describes when $L_c(\mathsf{triv})$ is finite dimensional, i.e. is supported at $0 \in \mathfrak{h}$. In the case c is a positive constant function, his result states that $x \in \mathfrak{h}$ is in the support of $L_c(\mathsf{triv})$ if and only if $P_W/P_{W_x}(e^{2\pi ic}) \neq 0$, where $P_W(q) = \sum_{w \in W} q^{\ell(w)}$ is the Poincaré polynomial of W.

The induction and restriction functors help to determine the set of aspherical values of W, [10].

Definition 7. The parameter $c \in \mathbb{C}[\mathcal{S}]^{\mathrm{ad}W}$ is an *aspherical value* of W if $eL_c(V) = 0$ for some $V \in \mathsf{Irrep}(W)$; such an $L_c(V)$ is called an *aspherical representation*. We let $\Sigma(W)$ denote the set of aspherical values of W.

It can be shown that $c \notin \Sigma(W)$ if and only if the functor $\mathsf{H}_{1,c}(W)$-mod $\longrightarrow$ $\mathsf{U}_{1,c}(W)$-mod, $M \mapsto eM$ is an equivalence. Thus for $c \notin \Sigma(W)$, $\mathsf{U}_{1,c}(W)$ inherits many favourable properties from $\mathsf{H}_{1,c}(W)$.

Using the restriction functors, one can show that $\Sigma(W)$ is the union of the $\Sigma(W')$ for proper parabolic subgroups $W' < W$ and of the set of finite dimensional aspherical representations of $\mathsf{H}_{1,c}(W)$. For $W = S_n$, this observation allows an inductive determination of the aspherical values, [10]. Remarkably, Bezrukavnikov and Etingof note that the number of aspherical representations matches phenomena in the $(\mathbb{C}^*)^2$-equivariant small quantum cohomology of $\mathsf{Hilb}^n(\mathbb{C}^2)$. Namely, multiplication in the quantum cohomology ring can be encoded by the so-called quantum differential equation which defines a flat connection on $\mathbb{C}$ for the trivial bundle associated with $H^*(\mathsf{Hilb}^n(\mathbb{C}^2), \mathbb{C})$, and this connection has regular singularities at $q = -\exp(2\pi ic)$ for $c \in \Sigma(S_n)$, [60]. Furthermore, the rank of the residue of the connection at each of these points equals the number of aspherical representations! For $W = G(d,1,n)$, the set $\Sigma(W)$ has been calculated by Dunkl and Griffeth, [23]; the quantum differential equation for $\mathsf{Hilb}^n(\widetilde{Y})$ of (1) has been described by Maulik and Oblomkov, [55]. A matching of data is again expected.

These surprising coincidences are part of a large programme involving several people which aims to study the quantum cohomology, and particularly the quantum differential equation, of symplectic resolutions of contracting symplectic singularities, [13]. Amongst other things, intriguing connections with geometric representation theory and with derived categories of symplectic resolutions are predicted, and representations of rational Cherednik algebras have an important role.

5. Reduction and Localisation

The spherical subalgebras $\mathsf{U}_{1,c}(W)$ share many properties with the quotients of enveloping algebras of reductive Lie algebras $U_\lambda(\mathfrak{g})$ at a central character λ. They are filtered with associated graded ring being the coordinate ring of a contracting symplectic singularity: $(\mathfrak{h} \times \mathfrak{h}^*)/W$ in the Cherednik case; $\mathcal{N}(\mathfrak{g})$ in the Lie case. This already produces a lot of structure including noetherianity, the Auslander-Gorenstein property, and a bound on the number of finite dimensional irreducible representations, [29]. Furthermore, it is only at very special values of the parameter where global dimension is infinite: at the aspherical values in the Cherednik case; at values such as $-\rho$ in the Lie case.

In the Lie case, a direct connection between $U_\lambda(\mathfrak{g})$ and the Springer resolution $\pi : T^*(G/B) \longrightarrow \mathcal{N}(\mathfrak{g})$ is made by the localisation theorem of Beilinson-Bernstein, [2]: this produces an equivalence between $U_\lambda(\mathfrak{g})$-modules and twisted $D_{G/B}$-modules. Combined with the Riemann-Hilbert correspondence, this relates BGG category $\mathcal{O}(\mathfrak{g})$ with perverse sheaves on G/B, and hence with Kazhdan-Lusztig theory for the Hecke algebra of the Weyl group of $\mathfrak{g}$.

We would like to produce an analogue of this for $\mathsf{U}_{1,c}(W)$ whenever there is a symplectic resolution $\pi : \widetilde{X} \longrightarrow (\mathfrak{h} \times \mathfrak{h}^*)/W$. This has been carried out for $W = S_n$ with $\widetilde{X} = \mathsf{Hilb}^n(\mathbb{C}^2)$, first in [45] algebraically, then in [34] and [50] using differential operators, then microlocal differential operators. (See [11] for similar results in positive characteristic.) Although these contructions are at their heart similar, and all have admitted various generalisations, the approaches in [34] and [50] connect directly to the mainstream of geometric representation theory. An interesting point is that, unlike $T^*(G/B)$, $\mathsf{Hilb}^n(\mathbb{C}^2)$ is not the cotangent bundle of a variety. This leads to a new point of view on localisation theorems which should be applicable to any symplectic resolution of a contracting symplectic singularity.

The first approach to quantising the Hilbert scheme follows Haiman's work on the $n!$ theorem, [47]. Here $\mathsf{Hilb}^n(\mathbb{C}^2)$ is constructed as the blow-up of $\mathsf{Sym}^n(\mathbb{C}^2)$ along the big diagonal, that is at the ideal $(\mathbb{C}[\mathfrak{h} \times \mathfrak{h}^*]^{\mathsf{sign}})^2$ where $\mathbb{C}[\mathfrak{h} \times \mathfrak{h}^*]^{\mathsf{sign}}$ denotes the polynomials that transform according to the sign representation under the S_n action. Thus $\operatorname{Coh} \mathsf{Hilb}^n(\mathbb{C}^2)$ is equivalent to a category of graded modules for the associated Rees ring. The first part of the following theorem asserts that there is a noncommutative version of this category.

Theorem 5.1 ([45]). *Assume that $c \not< 0$. There exists a category $\mathbb{X}_c$ of coherent sheaves on a noncommutative variety such that*

1. *$\mathbb{X}_c$ is a deformation of* $\operatorname{Coh} \mathsf{Hilb}^n(\mathbb{C}^2)$,
2. *There is an equivalence* $\mathsf{U}_{1,c}(S_n)$-mod $\xrightarrow{\sim} \mathbb{X}_c$.

The category $\mathbb{X}_c$ is a category of graded modules over an algebra which deforms the above Rees ring, replacing $\mathbb{C}[\mathfrak{h} \times \mathfrak{h}^*]^{S_n}$ with $\mathsf{U}_{1,c}(S_n)$ and $\mathbb{C}[\mathfrak{h} \times \mathfrak{h}^*]^{\mathsf{sign}}$

with $e\mathsf{H}_{1,c}(S_n)e_-$ where $e_- \in \mathbb{C}[S_n]$ is the idempotent corresponding to the sign representation. By an important result of Heckman-Opdam, see [9], $e\mathsf{H}_{1,c}(S_n)e_-$ is a $(\mathsf{U}_{1,c}(S_n), \mathsf{U}_{1,c+1}(S_n))$-bimodule and one can show it induces an equivalence $U_{1,c}(S_n)$-mod $\xrightarrow{\sim} U_{1,c+1}(S_n)$-mod whenever c and $c+1$ are not aspherical values. Thus the glueing data in the category $\mathbb{X}_c$ produces Morita equivalences, giving the second claim.

The advantage of this construction is that one can apply Haiman's work directly. This leads in [46] to the calculation of the characteristic cycle of any object from $\mathcal{O}_c(S_n)$, i.e. the support cycles in $\mathsf{Hilb}^n(\mathbb{C}^2)$ of the degeneration of the corresponding objects in $\mathbb{X}_c$; one can also show that the image in $\mathbb{X}_c$ of the $\mathsf{U}_{1,c}(S_n)$-module $e\mathsf{H}_{1,c}(S_n)$ is a deformation of the Procesi bundle $\mathcal{P}$ on $\mathsf{Hilb}^n(\mathbb{C}^2)$. In fact, since c is not aspherical $e\mathsf{H}_{1,c}(S_n)$ induces an equivalence between $\mathsf{U}_{1,c}(S_n)$-mod and $\mathsf{H}_{1,c}(S_n)$-mod and is thus a projective $\mathsf{U}_{1,c}(S_n)$-module carrying the regular representation of S_n. These properties are analogous to crucial properties of $\mathcal{P}$: it is an enduring hope that the representation theory of $\mathsf{H}_{1,c}(S_n)$ may be used to give a new proof of the $n!$ theorem.

Remarks 4. A similar algebraic analysis is carried out for kleinian singularities, [12] and [57], and for Cherednik algebras with $W = G(d,1,n)$, [41], but in this general case the geometry of the associated varieties generalising $\mathsf{Hilb}^n(\mathbb{C}^2)$ is not yet completely understood. There is also a localisation theorem for Harish-Chandra bimodules of finite W-algebras in this spirit, [36].

$\mathsf{Hilb}^n(\mathbb{C}^2)$ can be realised as a quiver variety, [58]. Let V be an n-dimensional vector space, and let $GL(V)$ act naturally on $Y = \mathsf{End}(V) \times V$. Set $X = T^*Y$ and let $\mu_X : X \to \mathfrak{gl}(V)^*$ be the moment map. Nakajima proved that the hamiltonian reduction $\mu_X^{-1}(0)//GL(V)$ is isomorphic to $\mathsf{Sym}^n(\mathbb{C}^2)$, and that there is an open set $X^s \subset X$ of "stable" representations on which $GL(V)$ acts freely such that $\mu_X^{-1}(0)^s/GL(V)$ is isomorphic to $\mathsf{Hilb}^n(\mathbb{C}^2)$ where $\mu_X^{-1}(0)^s := \mu_X^{-1}(0) \cap X^s$.

Differentiating the action of $GL(V)$ on Y produces a homomorphism $\tau_X : U(\mathfrak{gl}(V)) \longrightarrow D(Y)$, a noncommutative analogue of μ_X. If $\nu : \mathfrak{gl}(V) \longrightarrow \mathbb{C}$ is a character, let I_ν be the left ideal of $U(\mathfrak{gl}(V))$ generated by $A + \nu(A)$ for all $A \in \mathfrak{gl}(V)$ and let $(D_Y, GL(V))_\nu$-mod denote the category of $GL(V)$-equivariant D_Y-modules whose derived action of $\mathfrak{gl}(V)$ equals the action defined through $\tau_X + \nu$.

Theorem 5.2 ([34]). *Given a character $\nu : \mathfrak{gl}(V) \longrightarrow \mathbb{C}$, there is a parameter $c_\nu \in \mathbb{C}$ such that*

1. $(D(Y)/D(Y)\tau_X(I_\nu))^{GL(V)} \cong \mathsf{U}_{1,c_\nu}(S_n)$.
2. *There is a functor $\mathbb{H} : (D_Y, GL(V))_\nu$-mod $\longrightarrow \mathsf{U}_{1,c_\nu}(S_n)$-mod defined by $\mathbb{H}(M) = M^{GL(V)}$ which is exact and essentially surjective.*

The first part of this theorem quantises the quiver theoretic description of $\mathsf{Sym}^n(\mathbb{C}^2)$; the second part allows one to study $\mathsf{U}_{1,c}(S_n)$-modules via D-modules on Y.

To realise the Hilbert scheme instead, we must pass to the stable locus X^s. But D_Y-modules are local on the base Y rather than on $X = T^*Y$, and X^s is an open set defined on X. Thus we are led to a microlocal point of view, considering sheaves of algebras on X rather than on Y. There is a standard quantisation of the symplectic manifold $T^*\mathbb{C}^n$ via the Moyal product, producing a sheaf of $\mathbb{C}[[h]]$-algebras. Denote by $\mathcal{W}(T^*\mathbb{C}^n)$ the sheaf we get from this by inverting h. It is a sheaf of $\mathbb{C}(\!(h)\!)$-algebras.

Definition 8 ([50]). A *quantised differential operator algebra* on a smooth symplectic variety X is a sheaf of $\mathbb{C}(\!(h)\!)$-algebras, $\mathcal{W}_X$, such that for each $x \in X$ there is a neighbourhood U of x and a symplectic morphism $\phi : U \longrightarrow T^*\mathbb{C}^n$ such that $\mathcal{W}_X|_U \cong \phi^*\mathcal{W}(T^*\mathbb{C}^n)$.

Going back to our specific case let $U = X^s$, a symplectic manifold with a proper and free symplectic $GL(V)$-action and orbit map $p : \mu_X^{-1}(0)^s \longrightarrow \mu_X^{-1}(0)^s/GL(V) \cong \mathsf{Hilb}^n(\mathbb{C}^2)$. There is a noncommutative moment map: $\tau_U : \mathfrak{gl}(V) \longrightarrow \mathcal{W}_U$. Kashiwara and Rouquier, [50], show that

$$\mathcal{W}_{\mathsf{Hilb},\nu} := p_*\mathcal{E}nd_{\mathcal{W}}(\mathcal{W}_U/\mathcal{W}_U\tau_U(I_\nu))^{GL(V)}$$

is a quantised differential operator algebra on $\mathsf{Hilb}^n(\mathbb{C}^2)$ and that there is an equivalence of categories

$$(\mathcal{W}_{X^s}, GL(V))_\nu\text{-mod} \longrightarrow \mathcal{W}_{\mathsf{Hilb},\nu}\text{-mod}$$

for appropriate categories of $\mathcal{W}$-modules.

The categories above are $\mathbb{C}(\!(h)\!)$-linear and thus cannot be $D(Y)$-modules or $\mathsf{U}_{1,c}(S_n)$-modules. To remedy this, extend the good $\mathbb{C}^*$-actions that arise from the contracting action on $\mathsf{Sym}^n(\mathbb{C}^2)$ to the quantised differential operator algebras by letting h be an eigenvector of appropriate weight. Then categories of $\mathbb{C}^*$-equivariant $\mathcal{W}$-modules are equivalent, under taking fixed points, to $\mathbb{C}$-linear categories: for instance $(\mathcal{W}(T^*\mathbb{C}^n), \mathbb{C}^*)\text{-mod} \xrightarrow{\sim} D(\mathbb{C}^n)\text{-mod}$ for appropriate $\mathbb{C}^*$-actions. This produces an equivalence

$$(\mathcal{W}_{X^s}, GL(V) \times \mathbb{C}^*)_\nu\text{-mod} \longrightarrow (\mathcal{W}_{\mathsf{Hilb},\nu}, \mathbb{C}^*)\text{-mod},$$

the quantisation of the quiver theoretic description of $\mathsf{Hilb}^n(\mathbb{C}^2)$. Kashiwara and Rouquier then prove the following elegant Beilinson-Bernstein style theorem.

Theorem 5.3 ([50]). *For a character $\nu : \mathfrak{gl}(V) \longrightarrow \mathbb{C}$ such that $c_\nu \geq 0$, the global sections functor induces an equivalence*

$$(\mathcal{W}_{\mathsf{Hilb},\nu}, \mathbb{C}^*)\text{-mod} \longrightarrow \mathsf{U}_{1,c_\nu}(S_n)\text{-mod}\,.$$

With the approaches of [34] and of [50] one can begin a D-module or microlocal study of the representation theory of $\mathsf{U}_{1,c}(S_n)$ or $\mathsf{H}_{1,c}(S_n)$. This has been carried out (in a slightly different context) in [31] and [32]. Recently,

McGerty, [56], gives a new construction for $W = S_n$ of the KZ-functor, new versions of induction and restriction functors, and recovers the characteristic cycle computations of objects in $\mathcal{O}_c(S_n)$, all via microlocal fundamental groups and classical D-module theory from geometric representation theory.

Remarks 5. The above analysis should apply to other symplectic resolutions of contracting symplectic singularities that are realised by hamiltonian reduction. For finite W-algebras see [20] and for hypertoric varieties see [6] and the works of Braden, Licata, Proudfoot and Webster. For general quiver varieties one of the most intriguing aspects is to discover the algebras appearing as global sections, replacing the spherical Cherednik algebras in the Hilbert scheme case. It is still challenging to find the correct tools and concepts to unlock the properties of the categories of $\mathcal{W}$-modules.

Remarks 6. Back in the world of rational Cherednik algebras, it seems that the case $W = G(d, 1, n)$ will be understood via D-modules or microlocalisation. But, with the exception of G_4, all other complex reflection groups have no corresponding symplectic resolution; how to study these examples geometrically is unclear at the moment. That these cases have wider significance is clear from applications to integrable systems, D-module theory and the representation theory of complex reflection groups, see e.g. [8] and [7], and applications to algebraic combinatorics, see e.g. [40] and [43].

References

[1] A. Beauville, Symplectic singularities, *Invent. Math.* **139** (2000), 541–549.

[2] A. Beilinson and J. Bernstein, Localisation de $\mathfrak{g}$-modules, *C. R. Acad. Sci. Paris Sr. I Math.* **292** (1981), 15–18.

[3] G. Bellamy, On singular Calogero-Moser spaces, *Bull. Lond. Math. Soc.* **41** (2009), no. 2, 315–326.

[4] G. Bellamy, The Calogero-Moser partition for $G(m, d, n)$, *arXiv:0911.0066*.

[5] G. Bellamy, Cuspidal representations of rational Cherednik algebras at $t = 0$, *arXiv:0911.0069*.

[6] G. Bellamy and T. Kuwabara, On the deformation quantization of hypertoric varieties, *preprint*.

[7] Y. Berest and O. Chalykh, Quasi–invariants of complex reflection groups, *arXiv:0912.4518*.

[8] Y. Berest, P. Etingof and V. Ginzburg, Cherednik algebras and differential operators on quasi-invariants, *Duke Math. J.*, **118** (2003), 279–337.

[9] Y. Berest, P. Etingof and V. Ginzburg, Finite dimensional representations of rational Cherednik algebras, *Int. Math. Res. Not.*, **19** (2003), 1053–1088.

[10] R. Bezrukavnikov and P. Etingof, Parabolic induction and restriction functors for rational Cherednik algebras, *Selecta Math. (N.S.)* **14** (2009), no. 3–4, 397–425.

[11] R. Bezrukavnikov, M. Finkelberg and V. Ginzburg, Rational Cherednik algebras and Hilbert schemes in characteristic p, with an appendix by Pavel Etingof, *Represent. Theory* **10** (2006), 254–298.

[12] M. Boyarchenko, Quantization of minimal resolutions of Kleinian singularities, *Adv. Math.* **211** (2007), 244–265.

[13] A. Braverman, D. Maulik and A. Okounkov, Quantum cohomology of the Springer resolution, *arXiv:1001.0056*.

[14] M. Broué, G. Malle and R. Rouquier, Complex reflection groups, braid groups, Hecke algebras, *J. Reine Angew. Math.* **500** (1998), 127–190.

[15] K.A. Brown and K.R. Goodearl, Homological Aspects of Noetherian PI Hopf Algebras and Irreducible Modules of Maximal Dimension, *J. Algebra* **198** (1997), 240–265.

[16] K.A. Brown and I. Gordon, Poisson orders, representation theory and symplectic reflection algebras, *J. Reine Angew. Math.*, **559** (2003), 193–216.

[17] I. Burban and O. Schiffmann, On the Hall algebra of an elliptic curve, I, *arXiv:0505148*.

[18] I. Cherednik, *Double affine Hecke algebras*, London Mathematical Society Lecture Note Series, **319**, Cambridge University Press, Cambridge, (2005).

[19] W. Crawley-Boevey and M. Holland, Noncommutative deformations of Kleinian singularities, *Duke Math. J.* **92** (1998), 605–635.

[20] C. Dodd and K. Kremnizer, A localization theorem for finite W-algebras, *arXiv:0911.2210*.

[21] S. Donkin, *The q-Schur algebra*, London Mathematical Society Lecture Note Series, **253**, Cambridge University Press, Cambridge, (1998).

[22] C.F. Dunkl, Differential-difference operators associated to reflection groups, *Trans. Amer. Math. Soc.* **311** (1989), no. 1, 167–183.

[23] C.F. Dunkl and S. Griffeth, Generalized Jack polynomials and the representation theory of rational Cherednik algebras, *arXiv:1002.4607*.

[24] C.F.Dunkl and E.M. Opdam, Dunkl operators for complex reflection groups, *Proc. Lond. Math. Soc.*, **86** (2003) 70–108.

[25] P. Etingof, *Calogero-Moser systems and representation theory,* Zurich Lectures in Advanced Mathematics., European Mathematical Society, Zürich (2007).

[26] P. Etingof, Supports of irreducible spherical representations of rational Cherednik algebras of finite Coxeter groups, *arXiv:0911.3208*

[27] P. Etingof and V. Ginzburg, Symplectic reflection algebras, Calogero-Moser space, and deformed Harish-Chandra homomorphism, *Invent. Math.*, **147** (2002), 243–348.

[28] P. Etingof, X. Ma, Lecture notes on Cherednik algebras, *arXiv:1001.0432*.

[29] P. Etingof, T. Schedler and I. Losev Poisson traces and D-modules on Poisson varieties, *arXiv:0908.3868*.

[30] P. Etingof and E. Strickland, Lectures on quasi-invariants of Coxeter groups and the Cherednik algebra, *Enseign. Math.* **49** (2003), 35–65.

[31] M. Finkelberg and V. Ginzburg, Character sheaves and Cherednik algebras for algebraic curves, *arXiv:0704.3494*.

[32] M. Finkelberg, V. Ginzburg and R. Travkin, Mirabolic affine Grassmannian and character sheaves, *Selecta Math. (N.S.)* **14** (2009), no. 3–4, 607–628.

[33] B. Fu, A survey on symplectic singularities and resolutions, *Ann. Math. Blaise Pascal* **13** (2006), no. 2, 209–236.

[34] W.L. Gan and V. Ginzburg, Almost-commuting variety, D-modules, and Cherednik Algebras, with an appendix by Ginzburg, *IMRP Int. Math. Res. Pap.* (2006), 26439, 1–54.

[35] V. Ginzburg, On primitive ideals, *Selecta Math.*, **9** (2003), 379–407.

[36] V. Ginzburg, Harish-Chandra bimodules for quantized Slodowy slices., *Represent. Theory* **13** (2009), 236–271.

[37] V. Ginzburg, N. Guay, E. Opdam and R. Rouquier, On the category $\mathcal{O}$ for rational Cherednik algebras, *Invent. Math.*, **154** (2003), 617–651.

[38] V. Ginzburg and D. Kaledin, Poisson deformations of symplectic quotient singularities, *Adv. Math.* **186** (2004), 1–57.

[39] I. Gordon, Baby Verma modules for rational Cherednik algebras, *Bull. London Math. Soc.*, **35** (2003), 321–336.

[40] I. Gordon, On the quotient by diagonal invariants, *Invent. Math.*, **153** (2003), 503–518.

[41] I. Gordon, Quiver varieties, category O for rational Cherednik algebras, and Hecke algebras, *Int. Math. Res. Pap. IMRP* (2008), no. 3, Art. ID rpn006, 69 pp.

[42] I. Gordon, Symplectic reflection algebras, *Trends in representation theory of algebras and related topics*, 285–347, EMS Ser. Congr. Rep., Eur. Math. Soc., Zrich, 2008.

[43] I. Gordon and S. Griffeth, Catalan numbers for complex reflection groups, *arXiv:0912.1578*.

[44] I. Gordon and M. Martino, Calogero-Moser space, reduced rational Cherednik algebras, and two-sided cells, *Math. Res. Lett.* **16** (2009), no. 2, 255–262.

[45] I. Gordon and J.T. Stafford, Rational Cherednik algebras and Hilbert schemes I, *Adv. Math.* **198** (2005), 222–274.

[46] I. Gordon and J.T. Stafford, Rational Cherednik algebras and Hilbert schemes II: representations, *Duke Math. J.* **132** (2006), 73–135

[47] M. Haiman, Hilbert schemes, polygraphs, and the Macdonald positivity conjecture *J. Amer. Math. Soc.* **14** (2001), 941–1006.

[48] D. Kaledin, Symplectic singularities from the Poisson point of view, *J. Reine Angew. Math.*, **600** (2006), 135–156.

[49] D. Kaledin, Geometry and topology of symplectic resolutions., *Algebraic geometry—Seattle 2005*, Part 2, 595–628, Proc. Sympos. Pure Math., **80**, Part 2, Amer. Math. Soc., Providence, RI, 2009.

[50] M. Kashiwara and R. Rouquier, Microlocalization of rational Cherednik algebras, *Duke Math. J.* **144** (2008), no. 3, 525–573.

[51] L. Le Bruyn, Central singularities of quantum spaces., *J. Algebra* **177** (1995), no. 1, 142–153.

[52] M. Lehn and C. Sorger, A symplectic resolution for the binary tetrahedral group, *arXiv:0810.3225*.

[53] I. Losev, Completions of symplectic reflection algebras, *arXiv:1001.0239*.

[54] M. Martino,The Calogero-Moser partition and Rouquier families for complex reflection groups, *J. Algebra* **323** (2010), no. 1, 193–205

[55] D. Maulik, A. Oblomkov, Quantum cohomology of the Hilbert scheme of points on A_n-resolutions, *J. Amer. Math. Soc.* **22** (2009), no. 4, 1055–1091.

[56] K. McGerty, Microlocal KZ functors and rational Cherednik algebras, *preprint*.

[57] I. Musson, Hilbert schemes and noncommutative deformations of type A Kleinian singularities, *J. Algebra* **293** (2005), 102–129.

[58] H. Nakajima, *Lectures on Hilbert Schemes of Points on Surfaces*, Univ. Lecture Ser. **18**, Amer. Math. Soc., Providence, 1999.

[59] Y. Namikawa, Poisson deformations of affine symplectic varieties, *math.AG/ 0609741*

[60] A. Okounkov and R. Pandharipande, Quantum cohomology of the Hilbert scheme of points on the plane, *Invent. Math.* **179** (2010), 523–557.

[61] A. Premet, Special transverse slices and their enveloping algebras, *Adv. Math.* **170** (2002), no. 1, 1–55.

[62] R. Rouquier, Representations of rational Cherednik algebras, *Infinite-dimensional aspects of representation theory and applications,* Contemp. Math., **392**, Amer. Math. Soc., Providence, RI, (2005), 103–131.

[63] R. Rouquier, q-Schur algebras and complex reflection groups, *Mosc. Math. J.* **8** (2008), no. 1, 119–158.

[64] P. Shan, Crystals of Fock spaces and cyclotomic rational double affine Hecke algebras, *arXiv:0811.4549*

[65] G.C. Shephard and J.A. Todd, Finite unitary reflection groups, *Canadian J. Math.* **6** (1954), 274–304.

[66] T. Suzuki, Double affine Hecke algebras, conformal coinvariants and Kostka polynomials, *C. R. Math. Acad. Sci. Paris*, **343** (2006), 383–386.

[67] A. Tikaradze, On the Azumaya locus of an almost commutative algebra, *arXiv:0912.0307*.

[68] D. Uglov, Canonical bases of higher-level q-deformed Fock spaces and Kazhdan-Lusztig polynomials, in *Physical Combinatorics*, ed. M. Kashiwara, T. Miwa, Progress in Math. **191**, Birkhäuser (2000).

[69] M. Varagnolo and E. Vasserot,On the decomposition matrices of the quantized Schur algebra, *Duke Math. J.* **100** (1999), no. 2, 267–297.

[70] M. Varagnolo and E. Vasserot, Finite dimensional representations of DAHA and affine Springer fibers: the spherical case, *arXiv:0705.2691*.

[71] M. Varagnolo and E. Vasserot, Double affine Hecke algebras and affine flag manifolds, I, *arXiv:0911.5328*

[72] W. Wang, Hilbert schemes, wreath products, and the McKay correspondence, *arXiv:9912104*.

[73] X. Yvonne, A conjecture for q-decomposition matrices of cyclotomic v-Schur algebras, *J. Algebra* **304** (2006), 419–456.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Tensor Product Decomposition

Shrawan Kumar*

Abstract

Let G be a semisimple connected complex algebraic group. We study the tensor product decomposition of irreducible finite-dimensional representations of G. The techniques we employ range from representation theory to algebraic geometry and topology. This is mainly a survey of author's various results on the subject obtained individually or jointly with Belkale, Kapovich, Leeb, Millson and Stembridge.

Mathematics Subject Classification (2010). 20G05, 22E46

Keywords. Semisimple groups, tensor product decomposition, saturated tensor cone, PRVK conjecture, root components, geometric invariant theory.

Dedicated to the memory of my beloved mother

1. Introduction

Let G be a semisimple connected complex algebraic group with Lie algebra $\mathfrak{g}$. The irreducible finite-dimensional representations of G are parametrized by the set Λ^+ of dominant characters of T, where T is a maximal torus of G. For $\lambda \in \Lambda^+$, let $V(\lambda)$ be the corresponding (finite-dimensional) irreducible representation of G. By the complete reducibility theorem, for any $\lambda, \mu \in \Lambda^+$, we can decompose

$$V(\lambda) \otimes V(\mu) = \bigoplus_{\nu \in \Lambda^+} m^{\nu}_{\lambda,\mu} V(\nu), \tag{1}$$

where $m^{\nu}_{\lambda,\mu}$ (called the *Littlewood-Richardson coefficients*) denotes the multiplicity of $V(\nu)$ in the tensor product $V(\lambda) \otimes V(\mu)$. We say that $V(\nu)$ *occurs in* $V(\lambda) \otimes V(\mu)$ (or $V(\nu)$ is a *component* of $V(\lambda) \otimes V(\mu)$) if $m^{\nu}_{\lambda,\mu} > 0$. The numbers $m^{\nu}_{\lambda,\mu}$ are also called the *tensor product multiplicities*.

*The author was partially supported by NSF grants.
Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599–3250. E-mail: shrawan@email.unc.edu.

From the orthogonality relations, (1) is equivalent to the decomposition

$$\operatorname{ch} V(\lambda) \cdot \operatorname{ch} V(\mu) = \sum_{\nu \in \Lambda^+} m^{\nu}_{\lambda,\mu} \operatorname{ch} V(\nu). \tag{2}$$

One of the major goals of the 'tensor product problem' is to determine (all) the components of $V(\lambda) \otimes V(\mu)$. Of course, a more refined problem is to determine the components together with their multiplicities. In general, even the first problem is very hard.

We will also discuss a weaker 'saturated tensor product problem.' We say that $V(\nu)$ is a *saturated component* of $V(\lambda) \otimes V(\mu)$ if $V(N\nu)$ occurs in the tensor product $V(N\lambda) \otimes V(N\mu)$ for some integer $N \geq 1$.

The aim of this note is to give an overview of some of our results on the tensor product decomposition obtained individually or jointly with others over the last more than twenty years. We give enough details of many of the proofs to make this note more accessible.

We begin by setting the notation in Section 2 to be used through the paper. We recall some fairly well known basic facts (including some results of Kostant and Steinberg) about the tensor product decomposition in Section 3.

In Section 4, we recall the existence of 'root components' in the tensor product, conjectured by Wahl (and proved in [K_3]). Roughly, the result asserts that for any $\lambda, \mu \in \Lambda^+$ and any positive root β such that $\lambda + \mu - \beta \in \Lambda^+$, $V(\lambda + \mu - \beta)$ is a component of $V(\lambda) \otimes V(\mu)$ (cf. Theorem (4.1)). This result has a geometric counterpart in the surjectivity of the Wahl map for the flag varieties G/P (cf. Theorem (4.2)).

In Section 5, we study a solution of the Parthasarathy-Ranga Rao-Varadarajan-Kostant (for short PRVK) conjecture asserting that for $\lambda, \mu \in \Lambda^+$ and any $w \in W$, the irreducible G-module $V(\overline{\lambda + w\mu})$ occurs in the G-submodule $U(\mathfrak{g}) \cdot (v_\lambda \otimes v_{w\mu})$ of $V(\lambda) \otimes V(\mu)$ with multiplicity exactly 1, where W is the Weyl group of G, $\overline{\lambda + w\mu}$ denotes the unique element in Λ^+ in the W-orbit of $\lambda + w\mu$ and v_λ is a nonzero weight vector of $V(\lambda)$ of weight λ (cf. Theorem (5.13) and also its refinement Theorem (5.15)). We have outlined its more or less a complete proof except the proof of a crucial cohomology vanishing result for Bott-Samelson-Demazure-Hansen varieties (see Theorem (5.2)).

Section 6: This section is based on the work [BK_1] due to Belkale-Kumar. Since the existence of a component $V(\nu)$ in $V(\lambda) \otimes V(\mu)$ is equivalent to the nonvanishing of the G-invariant space $[V(\lambda) \otimes V(\mu) \otimes V(\nu^*)]^G$, the tensor product problem can be restated (replacing ν by ν^*) in a more symmetrical form of determining when $[V(\lambda) \otimes V(\mu) \otimes V(\nu)]^G \neq 0$. We generalize this problem from $s = 3$ to any $s \geq 1$ and define the *tensor product semigroup:*

$$\bar{\Gamma}_s(G) := \{(\lambda_1, \ldots, \lambda_s) \in (\Lambda^+)^s : [V(\lambda_1) \otimes \cdots \otimes V(\lambda_s)]^G \neq 0\}.$$

Similarly, define the *saturated tensor product semigroup*:

$$\Gamma_s(G) := \{(\lambda_1, \ldots, \lambda_s) \in (\Lambda^+)^s : [V(N\lambda_1) \otimes \cdots \otimes V(N\lambda_s)]^G \neq 0 \text{ for some } N > 0\}.$$

By virtue of the convexity result in symplectic geometry, there exists a (unique) convex polyhedral cone $\Gamma_s(G)_{\mathbb{R}} \subset (\Lambda^+_{\mathbb{R}})^s$ such that $\Gamma_s(G) = \Gamma_s(G)_{\mathbb{R}} \cap \Lambda^s$, where $\Lambda^+_{\mathbb{R}}$ is the dominant chamber in $\Lambda_{\mathbb{R}} := \Lambda \otimes_{\mathbb{Z}} \mathbb{R}$. The main result of this section (cf. Theorem (6.3)) determines a system of inequalities describing the cone $\Gamma_s(G)_{\mathbb{R}}$ explicitly in terms of a certain deformed product in the cohomology of the flag varieties G/P for maximal parabolic subgroups P. Moreover, as proved by Ressayre (cf. Theorem (6.4)), this system of inequalities is an irredundant system. We have outlined a more or less complete proof of Theorem (6.3), which makes essential use of Geometric Invariant Theory, specifically the Hilbert-Mumford criterion for semistability and Kempf's maximally destabilizing one parameter subgroups associated to unstable points. In addition, the notion of 'Levi-movability' plays a fundamental role in the proofs.

In Section 7, which is a joint work with Stembridge, we exploit isogenies between semisimple groups over algebraically closed fields of finite char. to get inequalities between the dimensions of invariants in tensor products of representations of complex semisimple groups (cf. Theorem (7.2)). As a corollary, we obtain that $\Gamma_s(\mathrm{Sp}(2\ell)) = \Gamma_s(\mathrm{SO}(2\ell+1))$ (cf. Corollary (7.5)).

Section 8 describes the 'saturation problem,' which provides a comparison between the semigroups $\Gamma_s(G)$ and $\bar{\Gamma}_s(G)$. We recall here the result due to Knutson-Tao on the saturation for the group $SL(n)$ and the results and conjectures of Kapovich-Millson and Belkale-Kumar.

Section 9 is devoted to recalling the classical Littlewood-Richardson theorem for the tensor product decomposition of irreducible polynomial representations of $\mathrm{GL}(n)$ and its generalization by Littelmann for any G via his *LS path model*. In addition, we recall the formula given by Berenstein-Zelevinsky, which determines the tensor product multiplicities as the number of lattice points in some convex polytope.

For the tensor product multiplicities, there is an approach by Lusztig [Lu] via his *canonical bases*. Similarly, there is an approach by Kashiwara [Ka] via his *crystal bases*.

There are some software programs to calculate the tensor product multiplicities (e.g., see [LCL], [St_1]). Also, for some explicit tensor product decompositions for $SL(n)$ see [BCH], [ST_2]; for E_8 see [MMP], [GP]; and for all the classical groups, see [Koi] and [L_1].

2. Notation

Let G be a semisimple connected complex algebraic group. We choose a Borel subgroup B and a maximal torus $T \subset B$ and let $W = W_G := N_G(T)/T$ be the associated Weyl group, where $N_G(T)$ is the normalizer of T in G. Let $P \supseteq B$ be a (standard) parabolic subgroup of G and let $U = U_P$ be its unipotent radical. Consider the Levi subgroup $L = L_P$ of P containing T, so that P is the semidirect product of U and L. Then, $B_L := B \cap L$ is a Borel subgroup of L. Let $\Lambda = \Lambda(T)$ denote the character group of T, i.e., the group of all the algebraic

group morphisms $T \to \mathbb{G}_m$. Clearly, W acts on Λ. We denote the Lie algebras of G, B, T, P, U, L, B_L by the corresponding Gothic characters: $\mathfrak{g}, \mathfrak{b}, \mathfrak{t}, \mathfrak{p}, \mathfrak{u}, \mathfrak{l}, \mathfrak{b}_L$ respectively. We will often identify an element λ of Λ (via its derivative $\dot{\lambda}$) by an element of $\mathfrak{t}^*$. Let $R = R_{\mathfrak{g}} \subset \mathfrak{t}^*$ be the set of roots of $\mathfrak{g}$ with respect to the Cartan subalgebra $\mathfrak{t}$ and let R^+ be the set of positive roots (i.e., the set of roots of $\mathfrak{b}$). Similarly, let $R_{\mathfrak{l}}$ be the set of roots of $\mathfrak{l}$ with respect to $\mathfrak{t}$ and $R_{\mathfrak{l}}^+$ be the set of roots of $\mathfrak{b}_L$. Let $\Delta = \{\alpha_1, \ldots, \alpha_\ell\} \subset R^+$ be the set of simple roots, $\{\alpha_1^\vee, \ldots, \alpha_\ell^\vee\} \subset \mathfrak{t}$ the corresponding simple coroots and $\{s_1, \ldots, s_\ell\} \subset W$ the corresponding simple reflections, where ℓ is the rank of G. We denote the corresponding simple root vectors by $\{e_1, \ldots, e_\ell\}$, i.e., $e_i \in \mathfrak{g}_{\alpha_i}$. We denote by $\Delta(P)$ the set of simple roots contained in $R_{\mathfrak{l}}$. For any $1 \leq j \leq \ell$, define the element $x_j \in \mathfrak{t}$ by

$$\alpha_i(x_j) = \delta_{i,j}, \ \forall\ 1 \leq i \leq \ell. \tag{3}$$

Recall that if W_P is the Weyl group of P (which is, by definition, the Weyl Group W_L of L), then in each coset of W/W_P we have a unique member w of minimal length. This satisfies (cf. [K$_4$, Exercise 1.3.E]):

$$wB_Lw^{-1} \subseteq B. \tag{4}$$

Let W^P be the set of the minimal length representatives in the cosets of W/W_P.

For any $w \in W^P$, define the Schubert cell:

$$C_w^P := BwP/P \subset G/P.$$

Then, it is a locally closed subvariety of G/P isomorphic with the affine space $\mathbb{A}^{\ell(w)}, \ell(w)$ being the length of w (cf. [J, Part II, Chapter 13]). Its closure is denoted by X_w^P, which is an irreducible (projective) subvariety of G/P of dimension $\ell(w)$. We denote the point $wP \in C_w^P$ by $\dot{w}$. We abbreviate X_w^B by X_w.

Let $\mu(X_w^P)$ denote the fundamental class of X_w^P considered as an element of the singular homology with integral coefficients $H_{2\ell(w)}(G/P, \mathbb{Z})$ of G/P. Then, from the Bruhat decomposition, the elements $\{\mu(X_w^P)\}_{w \in W^P}$ form a $\mathbb{Z}$-basis of $H_*(G/P, \mathbb{Z})$. Let $\{[X_w^P]\}_{w \in W^P}$ be the Poincaré dual basis of the singular cohomology with integral coefficients $H^*(G/P, \mathbb{Z})$. Thus, $[X_w^P] \in H^{2(\dim G/P - \ell(w))}(G/P, \mathbb{Z})$.

An element $\lambda \in \Lambda$ is called dominant (resp. dominant regular) if $\dot{\lambda}(\alpha_i^\vee) \geq 0$ (resp. $\dot{\lambda}(\alpha_i^\vee) > 0$) for all the simple coroots $\alpha_i^\vee$. Let Λ^+ (resp. Λ^{++}) denote the set of all the dominant (resp. dominant regular) characters. The set of isomorphism classes of irreducible (finite-dimensional) representations of G is parametrized by Λ^+ via the highest weight of an irreducible representation. For $\lambda \in \Lambda^+$, we denote by $V(\lambda)$ the corresponding irreducible representation (of highest weight λ). The dual representation $V(\lambda)^*$ is isomorphic with $V(\lambda^*)$, where λ^* is the weight $-w_o\lambda$; w_o being the longest element of W. The μ-weight space of $V(\lambda)$ is denoted by $V(\lambda)_\mu$. For $\lambda \in \Lambda^+$, let $P(\lambda)$ be the set of weights of $V(\lambda)$. The W-orbit of any $\lambda \in \Lambda$ contains a unique element in Λ^+, which we

denote by $\bar{\lambda}$. We also have the shifted action of W on Λ via $w*\lambda = w(\lambda+\rho)-\rho$, where ρ is half the sum of positive roots. (Observe that, even though ρ may not belong to Λ, $w\rho - \rho$ does.)

For any $\lambda \in \Lambda$, we have a G-equivariant line bundle $\mathcal{L}(\lambda)$ on G/B associated to the principal B-bundle $G \to G/B$ via the one-dimensional B-module λ^{-1}. (Any $\lambda \in \Lambda$ extends uniquely to a character of B.) The one-dimensional B-module λ is also denoted by $\mathbb{C}_\lambda$.

All the schemes are considered over the base field of complex numbers $\mathbb{C}$. The varieties are reduced (but not necessarily irreducible) schemes.

3. Some Basic Results

We follow the notation from the last section; in particular, G is a semisimple connected complex algebraic group. The aim of this section is to recall some fairly well known basic results on the tensor product decomposition. We begin with the following.

Lemma (3.1). *For $\lambda, \mu \in \Lambda^+$, $V(\lambda+\mu)$ occurs in $V = V(\lambda)\otimes V(\mu)$ with multiplicity 1.*

The unique submodule $V(\lambda+\mu)$ is called the Cartan component *of V.*

Proof. Let $v_\lambda \in V(\lambda)$ (resp. $v_\mu \in V(\mu)$) be a nonzero highest weight vector. Then, the line $\mathbb{C}v_\lambda \otimes v_\mu \subset V$ is clearly stable under the Borel subalgebra. From this, we easily see that the G-submodule generated by $v_\lambda \otimes v_\mu$ is isomorphic with $V(\lambda+\mu)$.

The weight space of V corresponding to the weight $\lambda+\mu$ is clearly one-dimensional. Hence, the multiplicity of $V(\lambda+\mu)$ in V is at most one. □

The following result is due to Kostant [Ko].

Proposition (3.2). *For $\lambda, \mu \in \Lambda^+$, any component $V(\nu)$ of $V = V(\lambda)\otimes V(\mu)$ is of the form $\nu = \lambda + \mu_1$, for some $\mu_1 \in P(\mu)$. Moreover, its multiplicity $m^\nu_{\lambda,\mu} \le \dim V(\mu)_{\mu_1}$.*

Proof. Clearly, the multiplicity $m^\nu_{\lambda,\mu}$ is equal to the dimension of

$$\begin{aligned}\operatorname{Hom}_{\mathfrak{g}}\big(V(\nu), V(\lambda)\otimes V(\mu)\big) &\simeq \operatorname{Hom}_{\mathfrak{b}}\big(\mathbb{C}_\nu, V(\lambda)\otimes V(\mu)\big)\\ &\simeq \operatorname{Hom}_{\mathfrak{b}}\big(\mathbb{C}_\nu \otimes V(\lambda)^*, V(\mu)\big).\end{aligned}$$

But, $V(\lambda)^*$ is generated, as a $\mathfrak{b}$-module, by its lowest weight vector $v_{-\lambda}$ of weight $-\lambda$. Hence, any homomorphism $\phi \in \operatorname{Hom}_{\mathfrak{b}}(\mathbb{C}_\nu \otimes V(\lambda)^*, V(\mu))$ is completely determined by $\phi(\mathbb{C}_\nu \otimes v_{-\lambda})$, which must be a weight vector of weight $-\lambda+\nu \in P(\mu)$. □

We have the following general result due to Steinberg [S].

Theorem (3.3). *For $\lambda, \mu, \nu \in \Lambda^+$, the multiplicity $m^{\nu}_{\lambda,\mu} = \sum_{w\in W} \varepsilon(w) n_{(w*\nu)-\lambda}(\mu)$, where $n_{\lambda'}(\mu)$ is the dimension of the λ'-weight space in $V(\mu)$.*

Proof. Define the $\mathbb{Z}$-linear operator $D : R(T) \to R(T)$ by $D(e^{\gamma}) = \frac{\sum_{w\in W} \varepsilon(w) e^{w*\gamma}}{\sum_{w\in W} \varepsilon(w) e^{w*0}}$, where $R(T)$ is the representation ring of the torus T. Then, D is linear over the invariant subring $R(T)^W$ (under the standard action of $W : v \cdot e^{\gamma} = e^{v\gamma}$). Moreover, $D(e^{v*\gamma}) = \varepsilon(v) D(e^{\gamma})$, for any $v \in W$. In particular, $D(e^{\gamma}) = 0$ if $\gamma + \rho$ is not regular (equivalently, if γ has nontrivial isotropy under the shifted action of W). For any $\gamma \in \Lambda$ such that $\gamma + \rho$ is regular, let $w_{\gamma} \in W$ be the unique element such that $w_{\gamma} * \gamma \in \Lambda^+$.

By the Weyl character formula, for any $\lambda \in \Lambda^+$, $\operatorname{ch} V(\lambda) = D(e^{\lambda})$, where $\operatorname{ch} V(\lambda)$ denotes the character of $V(\lambda)$. Thus,

$$\begin{aligned}
\operatorname{ch}(V(\lambda) \otimes V(\mu)) &= \operatorname{ch} V(\lambda) \cdot \operatorname{ch} V(\mu) \\
&= D(e^{\lambda} \cdot \operatorname{ch} V(\mu)), \qquad \text{since } \operatorname{ch} V(\mu) \in R(T)^W \\
&= \sum_{\gamma} n_{\gamma}(\mu) D(e^{\lambda} \cdot e^{\gamma}) \\
&= \sum_{\substack{\gamma : \lambda+\gamma+\rho \\ \text{is regular}}} \varepsilon(w_{\lambda+\gamma})\, n_{\gamma}(\mu)\, D\Big(e^{w_{\lambda+\gamma^*}(\lambda+\gamma)}\Big) \\
&= \sum_{\nu\in\Lambda^+} \left(\sum_{w\in W} \varepsilon(w)\, n_{(w*\nu)-\lambda}(\mu) \right) D(e^{\nu}), \text{ since } \varepsilon(w) = \varepsilon(w^{-1}).
\end{aligned}$$

Thus, from the equivalence of (1) and (2) in Section 1, the theorem follows. □

Corollary (3.4). *For $\lambda, \mu \in \Lambda^+$, if $\lambda + \mu'$ is nearly dominant (i.e., $(\lambda + \mu')(\alpha_i^{\vee}) \geq -1$ for all the simple coroots $\alpha_i^{\vee}$) for all μ' in $P(\mu)$, then the multiplicity of $V(\nu)$ in $V(\lambda) \otimes V(\mu)$:*

$$m^{\nu}_{\lambda,\mu} = n_{\nu-\lambda}(\mu).$$

Of course, by Proposition (3.2), $V(\nu)$ occurs in $V(\lambda) \otimes V(\mu)$ only if $\nu = \lambda + \mu'$ for some $\mu' \in P(\mu)$.

Proof. By the above theorem,

$$m^{\nu}_{\lambda,\mu} = \sum_{w\in W} \varepsilon(w)\, n_{(w*\nu)-\lambda}(\mu).$$

For $w \neq 1$, we claim that $n_{(w*\nu)-\lambda}(\mu) = 0$. Equivalently, $(w*\nu) - \lambda \notin P(\mu)$. Since any weight in $\lambda + P(\mu)$ is nearly dominant (by assumption) and ν is dominant, we have $w * \nu \notin \lambda + P(\mu)$ for any $w \neq 1$. □

As a corollary of the above corollary, we get the following (cf. [Kas], [K_2, Proposition 1.5]).

Corollary (3.5). *For $\lambda, \mu \in \Lambda^+$ such that $V(\mu)$ is minuscule (i.e., $P(\mu)$ is a single W-orbit), we have the decomposition*

$$V(\lambda) \otimes V(\mu) \simeq \bigoplus_{\substack{\bar{w} \in W/W_\mu: \\ \lambda + w\mu \in \Lambda^+}} V(\lambda + w\mu), \tag{$*$}$$

each occuring with multiplicity 1, where $W_\mu := \{w \in W : w\mu = \mu\}$ is the isotropy group of μ. Moreover, the number of irreducible components in $V(\lambda) \otimes V(\mu)$ is equal to the cardinality $\#W_\lambda \backslash W/W_\mu$.

Proof. By [Bo, Exercise 24, p. 226], $\lambda + \mu'$ is nearly dominant for any $\mu' \in P(\mu)$. Thus, by the above corollary, the decomposition $(*)$ follows. For the second part, define

$$f : (W/W_\mu)^+ \to W_\lambda \backslash W/W_\mu, \quad f(wW_\mu) = W_\lambda w W_\mu,$$

where $(W/W_\mu)^+ := \{\bar{w} \in W/W_\mu : \lambda + w\mu \in \Lambda^+\}$. It is easy to see that f is injective, and, for any w of minimal element in its double coset $W_\lambda w W_\mu$, $wW_\mu \in (W/W_\mu)^+$. □

As another corollary of Theorem (3.3), we get the following.

Corollary (3.6). *For $\lambda, \mu, \nu \in \Lambda^+$,*

$$m^\nu_{\lambda,\mu} = \sum_{v,w \in W} \varepsilon(v)\, \varepsilon(w)\, \mathcal{P}\big(v(\mu + \rho) - w(\nu + \rho) + \lambda\big),$$

where $\mathcal{P}$ is the Kostant's partition function.

Proof. Use Kostant's formula for any dominant character μ and any integral character λ':

$$n_{\lambda'}(\mu) = \sum_{v \in W} \varepsilon(v)\, \mathcal{P}((v * \mu) - \lambda').$$

□

The following result is due to Kostant [Ko, Lemma 4.1].

Theorem (3.7). *For any $\lambda, \mu, \nu \in \Lambda^+$, the multiplicity*

$$m^\nu_{\lambda,\mu} = \dim\Big\{v \in V(\mu)_{\nu - \lambda} : e_i^{\lambda(\alpha_i^\vee)+1} v = 0, \text{ for all simple roots } \alpha_i\Big\}.$$

Proof. Of course, by the proof of Proposition (3.2),

$$\begin{aligned} m^\nu_{\lambda,\mu} &= \dim \mathrm{Hom}_{\mathfrak{g}}(V(\nu), V(\lambda) \otimes V(\mu)) \\ &= \dim \mathrm{Hom}_{\mathfrak{b}}(\mathbb{C}_\nu \otimes V(\lambda)^*, V(\mu)). \end{aligned}$$

Let $v_{-\lambda} \in V(\lambda)^*$ be a nonzero lowest weight vector. Then, by a result due to Harish-Chandra,

$$\phi : U(\mathfrak{n}) \longrightarrow V(\lambda)^*, \quad X \mapsto X \cdot v_{-\lambda},$$

is surjective with kernel

$$\ker \phi = \sum_{1 \le i \le \ell} U(\mathfrak{n}) \cdot e_i^{\lambda(\alpha_i^\vee)+1},$$

where $\mathfrak{n}$ is the nil-radical of $\mathfrak{b}$. (This also follows immediately from the BGG resolution.) This proves the theorem. □

The following corollary follows immediately from the above theorem and SL(2)-representation theory.

Corollary (3.8). *For any $\lambda, \mu \in \Lambda^+$ and $w \in W$ such that $\lambda + w\mu$ is dominant, we have $m_{\lambda,\mu}^{\lambda+w\mu} = 1$.*

Lemma (3.9). *For any $\lambda, \mu, \nu, \lambda', \mu', \nu' \in \Lambda^+$ such that $m_{\lambda',\mu'}^{\nu'} \ge 1$, we have*

$$m_{\lambda+\lambda',\mu+\mu'}^{\nu+\nu'} \ge m_{\lambda,\mu}^{\nu}.$$

Proof. We have

$$\begin{aligned} \mathrm{Hom}_{\mathfrak{g}}\big(V(\nu), V(\lambda) \otimes V(\mu)\big) &\simeq \mathrm{Hom}_{\mathfrak{g}}\big(V(\lambda)^* \otimes V(\mu)^* \otimes V(\nu^*)^*, \mathbb{C}\big) \\ &\simeq \big[V(\lambda)^* \otimes V(\mu)^* \otimes V(\nu^*)^*\big]^{\mathfrak{g}} \\ &\simeq H^0\big((G/B)^3, \mathcal{L}(\lambda \boxtimes \mu \boxtimes \nu^*)\big)^G, \end{aligned}$$

where the last isomorphism follows from the Borel-Weil theorem: $H^0(G/B, \mathcal{L}(\lambda)) \simeq V(\lambda)^*$ (for any $\lambda \in \Lambda^+$), and $\mathcal{L}(\lambda \boxtimes \mu \boxtimes \nu^*)$ denotes the external tensor product line bundle $\mathcal{L}(\lambda) \boxtimes \mathcal{L}(\mu) \boxtimes \mathcal{L}(\nu^*)$ on $(G/B)^3$. Take a nonzero $\sigma_o \in H^0\big((G/B)^3, \mathcal{L}(\lambda' \boxtimes \mu' \boxtimes \nu'^*)\big)^G$. Then, the map

$$H^0\big((G/B)^3, \mathcal{L}(\lambda \boxtimes \mu \boxtimes \nu^*)\big)^G \longrightarrow H^0\big((G/B)^3, \mathcal{L}\big((\lambda+\lambda') \boxtimes (\mu+\mu') \boxtimes (\nu^*+\nu'^*)\big)\big)^G,$$

$\sigma \mapsto \sigma \cdot \sigma_o$, is clearly injective. □

4. Root Components in the Tensor Product

In this section, we assume that G is a semisimple simply-connected complex algebraic group and follow the notation from Section 2. The aim of this section is to state the existence of certain tensor product components coming from the positive roots, conjectured by Wahl [W]. Specifically, we have the following result due to Kumar [K$_3$], a proof of which can be found in loc. cit. The proof is purely representation theoretic (based on Theorem (3.7)) and unfortunately requires some case by case analysis. For any $\lambda \in \Lambda$, define $S_\lambda = \{1 \le i \le \ell : \lambda(\alpha_i^\vee) = 0\}$. Also, for any $\beta \in R^+$, define $F_\beta = \{1 \le i \le \ell : \beta - \alpha_i \notin R^+ \cup \{0\}\}$.

Theorem (4.1). *Take any $\lambda, \mu \in \Lambda^+$ and $\beta \in R^+$ satisfying:*

(P_1) $\lambda + \mu - \beta \in \Lambda^+$, and

(P_2) $S_\lambda \cup S_\mu \subset F_\beta$.

Then, $V(\lambda + \mu - \beta)$ is a component of $V(\lambda) \otimes V(\mu)$.

Observe that if G_2 does not occur as a component of $\mathfrak{g}$, then the conditions $(P_1) - (P_2)$ are automatically satisfied for any $\lambda, \mu \in \Lambda^{++}$.

Let X be a smooth projective variety with line bundles $\mathcal{L}_1$ and $\mathcal{L}_2$ on X. Consider the *Wahl map* defined by him (which he called the *Gaussian map*) $\Phi_{\mathcal{L}_1,\mathcal{L}_2} : H^0(X \times X, \mathcal{I}_D \otimes (\mathcal{L}_1 \boxtimes \mathcal{L}_2)) \to H^0(X, \Omega^1_X \otimes \mathcal{L}_1 \otimes \mathcal{L}_2)$, which is induced from the projection $\mathcal{I}_D \to \mathcal{I}_D/\mathcal{I}_D^2$ by identifying the $\mathcal{O}_{X\times X}/\mathcal{I}_D \simeq \mathcal{O}_D$-module $\mathcal{I}_D/\mathcal{I}_D^2$ (supported in D) with the sheaf of 1-forms Ω^1_X on $D \simeq X$ (cf. [W]), where $\mathcal{I}_D$ is the ideal sheaf of the diagonal D.

The following Theorem is a geometric counterpart of Theorem (4.1). It was conjectured by Wahl and proved by him for $X = \mathrm{SL}(n)/B$ and also for any minuscule G/P (cf. [W]). Kumar proved it for any G/P (cf. [K_3]) by using Theorem (4.1). In fact, he showed that Theorems (4.1) and (4.2) are 'essentially' equivalent. Theorem (4.2) is proved in an arbitrary char. for Grassmannians by Mehta-Parameswaran [MP]; for orthogonal and symplectic Grassmannians in odd char. by Lakshmibai-Raghavan-Sankaran [LRS]; and for minuscule G/P in any char. by Brown-Lakshmibai [BL].

Theorem (4.2). *The Wahl map $\Phi_{\mathcal{L}_1,\mathcal{L}_2}$ is surjective for any flag variety $X = G/P$ (where G is any semisimple simply-connected group and $P \subset G$ a parabolic subgroup) and any ample homogeneous line bundles $\mathcal{L}_1$ and $\mathcal{L}_2$ on X.*

Equivalently, $H^p(G/P \times G/P, \mathcal{I}_D^2 \otimes (\mathcal{L}_1 \boxtimes \mathcal{L}_2)) = 0$, for all $p > 0$.

5. Proof of Parthasarathy-Ranga Rao-Varadarajan-Kostant Conjecture

In this section, we assume that G is a semisimple simply-connected complex algebraic group and follow the notation from Section 2. We begin with the following result due to Parthasarathy-Ranga Rao-Varadarajan [PRV, Corollary 1 to Theorem 2.1].

Theorem (5.1). *For any $\lambda, \mu \in \Lambda^+$, the irreducible module $V(\overline{\lambda + w_o\mu})$ occurs with multiplicity one in the tensor product $V(\lambda) \otimes V(\mu)$, where w_o is the longest element of W.*

Proof. Denote $\nu = \overline{\lambda + w_o\mu}$. We clearly have

$$\mathrm{Hom}_{\mathfrak{g}}(V(\lambda) \otimes V(\mu), V(\nu)) \simeq \mathrm{Hom}_{\mathfrak{b}}(\mathbb{C}_\lambda \otimes V(\mu), V(\nu)).$$

Moreover, as in the proof of Theorem (3.7), the map $\phi : U(\mathfrak{n}) \to V(\mu), X \mapsto X \cdot v_{w_o\mu}$, is surjective with kernel

$$\ker \phi = \sum_{1 \le i \le \ell} U(\mathfrak{n}) e_i^{-(w_o\mu)(\alpha_i^\vee)+1}, \tag{5}$$

where $v_{w_o\mu}$ is a nonzero lowest weight vector of $V(\mu)$. Since the weight space $V(\nu)_{\lambda+w_o\mu}$ is one-dimensional, $\dim \mathrm{Hom}_{\mathfrak{b}}(\mathbb{C}_\lambda \otimes V(\mu), V(\nu)) \le 1$. Moreover, by (5), the map $v_\lambda \otimes v_{w_o\mu} \mapsto v_{\lambda+w_o\mu}$ extends to a $\mathfrak{b}$-module map $\mathbb{C}_\lambda \otimes V(\mu) \to V(\nu)$ iff $e_i^{-(w_o\mu)(\alpha_i^\vee)+1} v_{\lambda+w_o\mu} = 0$ for all $1 \le i \le \ell$. But the latter holds, as can be easily seen from the representation theory of SL_2. □

Now, we prove a vast generalization of the above theorem.

For any B-variety X, we denote by $\widetilde{X}$ the G-variety $G \underset{B}{\times} X$, i.e., it is the total space of the fiber bundle with fiber X, associated to the principal B-bundle $G \to G/B$. For any B-varieties X, Y and a B-morphism $\phi : X \to Y$, there is a canonical G-morphism $\widetilde{\phi} : \widetilde{X} \to \widetilde{Y}$.

For any sequence (not necessarily reduced) $\mathfrak{w} = (s_{i_1}, \dots, s_{i_n})$ of simple reflections, let $Z_{\mathfrak{w}}$ be the Bott-Samelson-Demazure-Hansen (for short BSDH) variety defined as the quotient $Z_{\mathfrak{w}} = P_{i_1} \times \cdots \times P_{i_n}/B^n$ under the right action of B^n on $P_{i_1} \times \cdots \times P_{i_n}$ via:

$$(p_1, \dots, p_n)(b_1, \dots, b_n) = (p_1 b_1, b_1^{-1} p_2 b_2, \dots, b_{n-1}^{-1} p_n b_n),$$

for $p_j \in P_{i_j}, b_j \in B$, where P_{i_j} is the standard minimal parabolic with $\Delta(P_{i_j}) = \{\alpha_{i_j}\}$. We denote the B^n-orbit of $(p_1, \dots, p_n)$ by $[p_1, \dots, p_n]$. Then, $Z_{\mathfrak{w}}$ is a smooth B-variety (in fact a P_{i_1}-variety) under the left multiplication on the first factor. For any $1 \le j \le n$, consider the subsequence $\mathfrak{w}(j) := (s_{i_1}, \dots, \hat{s}_{i_j}, \dots, s_{i_n})$. Then, we have a B-equivariant embeding $Z_{\mathfrak{w}(j)} \hookrightarrow Z_{\mathfrak{w}}, [p_1, \dots, \hat{p}_j, \dots, p_n] \mapsto [p_1, \dots, p_{j-1}, 1, p_{j+1}, \dots, p_n]$. Thus, we have the G-varieties $\widetilde{Z}_{\mathfrak{w}}$ and $\widetilde{Z}_{\mathfrak{w}(j)}$ and a canonical inclusion $\widetilde{Z}_{\mathfrak{w}(j)} \hookrightarrow \widetilde{Z}_{\mathfrak{w}}$. Of course, $\widetilde{Z}_{\mathfrak{w}}$ (and $\widetilde{Z}_{\mathfrak{w}(j)}$) is smooth. For any $w \in W$, we also have the G-variety $\widetilde{X}_w$, where X_w is the Schubert variety as in Section 2. Moreover, for any $v \le w$, we have a canonical inclusion $\widetilde{X}_v \hookrightarrow \widetilde{X}_w$, induced from the inclusion $X_v \hookrightarrow X_w$. Further, there are G-morphisms (G acting on $G/B \times G/B$ diagonally):

$$\widetilde{\theta}_{\mathfrak{w}} : \widetilde{Z}_{\mathfrak{w}} \to G/B \times G/B \text{ and } \widetilde{d}_w : \widetilde{X}_w \to G/B \times G/B,$$

defined by

$$\widetilde{\theta}_{\mathfrak{w}}[g, z] = (gB, g\theta_{\mathfrak{w}}(z)), \text{ for } g \in G, z \in Z_{\mathfrak{w}}, \text{ and}$$

$$\widetilde{d}_w[g, x] = (gB, gx), \text{ for } g \in G, x \in X_w,$$

where the map $\theta_{\mathfrak{w}} : Z_{\mathfrak{w}} \to G/B$ is defined by $[p_1, \dots, p_n] \mapsto p_1 \dots p_n B$. Clearly, the map $\widetilde{\theta}_{\mathfrak{w}}$ (resp. $\widetilde{d}_w$) is well defined, i.e., it descends to $\widetilde{Z}_{\mathfrak{w}}$ (resp. $\widetilde{X}_w$). It can be easily seen that the map $\widetilde{d}_w$ is a closed immersion and its image is

the closure of the G-orbit of the point $(e, \dot{w})$ in $G/B \times G/B$, where $\dot{w}$ is the point $wB \in G/B$. The sequence $\mathfrak{w} = (s_{i_1}, \dots, s_{i_n})$ is said to be *reduced* if $m(\mathfrak{w}) := s_{i_1} \dots s_{i_n}$ is a reduced decomposition.

For any $\lambda, \mu \in \Lambda$, we denote by $\mathcal{L}(\lambda \boxtimes \mu)$ the line bundle on $G/B \times G/B$ which is the external tensor product of the line bundles $\mathcal{L}(\lambda)$ and $\mathcal{L}(\mu)$ respectively. We further denote by $\mathcal{L}_{\mathfrak{w}}(\lambda \boxtimes \mu)$ (resp. $\mathcal{L}_w(\lambda \boxtimes \mu)$) the pull-back of $\mathcal{L}(\lambda \boxtimes \mu)$ by the map $\widetilde{\theta}_{\mathfrak{w}}$ (resp. $\widetilde{d}_w$). The following cohomology vanishing result (rather its Corollary (5.4)) is crucial to the proof of the PRVK conjecture.

Theorem (5.2). *Let $\mathfrak{w} = (s_{i_1}, \dots, s_{i_n})$ be any sequence of simple reflections and let $1 \le j \le k \le n$ be such that the subsequence $(s_{i_j}, \dots, s_{i_k})$ is reduced. Then, for any $\lambda, \mu \in \Lambda^+$, we have:*

$$H^p\Big(\widetilde{Z}_{\mathfrak{w}}, \mathcal{L}_{\mathfrak{w}}(\lambda \boxtimes \mu) \otimes \mathcal{O}_{\widetilde{Z}_{\mathfrak{w}}}\big[-\cup_{q=j}^{k} \widetilde{Z}_{\mathfrak{w}(q)}\big]\Big) = 0, \text{ for all } p > 0.$$

The proof of this theorem is identical to the proof of the analogous result for $Z_{\mathfrak{w}}$ given in [K$_4$, Theorem 8.1.8] if we observe the following simple:

Lemma (5.3). *For any sequence $\mathfrak{w} = (s_{i_1}, \dots, s_{i_n})$ (not necessarily reduced), the canonical bundle $K_{\widetilde{Z}_{\mathfrak{w}}}$ of $\widetilde{Z}_{\mathfrak{w}}$ is isomorphic with*

$$\mathcal{L}_{\mathfrak{w}}((-\rho) \boxtimes (-\rho)) \otimes \mathcal{O}_{\widetilde{Z}_{\mathfrak{w}}}\big[-\partial \widetilde{Z}_{\mathfrak{w}}\big], \text{ where } \partial \widetilde{Z}_{\mathfrak{w}} := \cup_{q=1}^{n} \widetilde{Z}_{\mathfrak{w}(q)}.$$

Applying Theorem (5.2) to the cohomology exact sequence, corresponding to the sheaf sequence:

$$0 \to \mathcal{O}_{\widetilde{Z}_{\mathfrak{w}}}[-\widetilde{Z}_{\mathfrak{w}(j)}] \to \mathcal{O}_{\widetilde{Z}_{\mathfrak{w}}} \to \mathcal{O}_{\widetilde{Z}_{\mathfrak{w}(j)}} \to 0$$

tensored with the locally free sheaf $\mathcal{L}_{\mathfrak{w}}(\lambda \boxtimes \mu)$, we get the following:

Corollary (5.4). *Let $\mathfrak{w} = (s_{i_1}, \dots, s_{i_n})$ be any sequence. Then, for any $1 \le j \le n$ and any $\lambda, \mu \in \Lambda^+$, the canonical restriction map $H^0\big(\widetilde{Z}_{\mathfrak{w}}, \mathcal{L}_{\mathfrak{w}}(\lambda \boxtimes \mu)\big) \to H^0\big(\widetilde{Z}_{\mathfrak{w}(j)}, \mathcal{L}_{\mathfrak{w}(j)}(\lambda \boxtimes \mu)\big)$ is surjective.*

In the case when $\mathfrak{w}$ is a reduced sequence, the image of the map $\widetilde{\theta}_{\mathfrak{w}} : \widetilde{Z}_{\mathfrak{w}} \to G/B \times G/B$ is precisely equal to $\widetilde{d}_w(\widetilde{X}_w)$, where $w = m(\mathfrak{w})$. By slight abuse of notation, we denote the map $\widetilde{\theta}_{\mathfrak{w}}$, considered as a map $\widetilde{Z}_{\mathfrak{w}} \to \widetilde{X}_w$, again by $\widetilde{\theta}_{\mathfrak{w}}$. Then, $\widetilde{\theta}_{\mathfrak{w}}$ is a birational surjective morphism. As a consequence of the above corollary, we get the following:

Corollary (5.5). *For any $v \le w \in W$, and $\lambda, \mu \in \Lambda^+$, the canonical restriction map $H^0\big(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu)\big) \to H^0\big(\widetilde{X}_v, \mathcal{L}_v(\lambda \boxtimes \mu)\big)$ is surjective.*

Proof. Take a reduced sequence $\mathfrak{w}$ such that $m(\mathfrak{w}) = w$. Then, we can find a reduced subsequence $\mathfrak{v}$ such that $m(\mathfrak{v}) = v$ (cf. [K$_4$, Lemma 1.3.16]). Since any Schubert variety X_w is normal (cf. [BrK, Theorem 3.2.2]), then so is $\widetilde{X}_w$. Hence,

by the projection formula [H, Exercise 5.1, Chap. II] and the Zariski's main theorem [H, Corollary 11.4 and its proof, Chap. III] applied to the projective birational morphism $\widetilde{\theta}_{\mathfrak{w}}$, we get the isomorphism

$$\widetilde{\theta}^*_{\mathfrak{w}} : H^0\big(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu)\big) \simeq H^0\big(\widetilde{Z}_{\mathfrak{w}}, \mathcal{L}_{\mathfrak{w}}(\lambda \boxtimes \mu)\big).$$

Now, the corollary follows by successively applying the last corollary. □

Remark (5.6). Even though we do not need, we also get (from Theorem (5.2)) that for any locally free sheaf $\mathcal{L}$ on $\widetilde{X}_w$, one has:

$$H^p(\widetilde{X}_w, \mathcal{L}) \simeq H^p(\widetilde{Z}_{\mathfrak{w}}, \widetilde{\theta}^*_{\mathfrak{w}}(\mathcal{L})), \text{ for all } p \geq 0,$$

and $H^p\big(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu)\big) = 0$ for all $p > 0$ and any $\lambda, \mu \in \Lambda^+$. These cohomological results hold even in an arbitrary char. via Frobenius splitting methods (cf. [BrK, Theorems 3.1.2 and 3.3.4]).

The following result is a special case of a theorem of Bott [Bot, Theorem I], who proved the result for an arbitrary $H^p(G/B, \mathcal{M})$ in terms of the Lie algebra cohomology (cf. [K$_4$, Exercise 8.3.E.4] for the statement and the idea of a short proof).

Theorem (5.7). *For any finite-dimensional algebraic B-module M, there is a G-module isomorphism:*

$$H^0(G/B, \mathcal{M}) \simeq \bigoplus_{\theta \in \Lambda^+} V(\theta)^* \otimes [V(\theta) \otimes M]^{\mathfrak{b}},$$

where we put the trivial G-module structure on the $\mathfrak{b}$-invariants and $\mathcal{M}$ denotes the locally free sheaf on G/B associated to the B-module M.

Proof. By the Peter-Weyl theorem and Tannaka-Krein duality (cf. [BD, Chap. III]), the affine coordinate ring $\mathbb{C}[G]$ (as a $G \times G$-module) is given by:

$$\mathbb{C}[G] \simeq \bigoplus_{\theta \in \Lambda^+} V(\theta)^* \otimes V(\theta),$$

where $G \times G$ acts on $\mathbb{C}[G]$ via $\big((g,h).f\big)(x) = f(g^{-1}xh)$ and $G \times G$ acts on $V(\theta)^* \otimes V(\theta)$ factorwise. From this, the theorem follows easily. □

As a consequence of the above theorem, we derive the following:

Theorem (5.8). *For any $w \in W, \lambda \in \Lambda$ and $\mu \in \Lambda^+$, $H^0\big(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu)\big)$ is canonically G-module isomorphic with*

$$\bigoplus_{\theta \in \Lambda^+} V(\theta)^* \otimes \mathrm{Hom}_{\mathfrak{b}}(\mathbb{C}_\lambda \otimes V_w(\mu), V(\theta)),$$

where we put the trivial G-module structure on $\mathrm{Hom}_{\mathfrak{b}}(\mathbb{C}_\lambda \otimes V_w(\mu), V(\theta))$ and $V_w(\mu) \subset V(\mu)$ is the Demazure submodule, which is, by definition, the $U(\mathfrak{b})$-span of the extremal weight vector $v_{w\mu}$ of weight $w\mu$ in $V(\mu)$.

Proof. By the definition of the direct image sheaf π_*, corresponding to the canonical fibration $\pi = \pi_w : \widetilde{X}_w \to G/B$, we get that $H^0\big(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu)\big) \simeq H^0\big(G/B, \pi_*\mathcal{L}_w(\lambda \boxtimes \mu)\big)$. Since the line bundle $\mathcal{L}_w(\lambda \boxtimes \mu)$ on the G-space $\widetilde{X}_w$ is a G-equivariant line bundle and the map π is G-equivariant, the direct image sheaf $\pi_*\mathcal{L}_w(\lambda \boxtimes \mu)$ is a locally free sheaf on G/B associated to the B-module $M_w := \mathbb{C}_{-\lambda} \otimes H^0(X_w, \mathcal{L}_w(\mu))$, where $\mathcal{L}_w(\mu) := \mathcal{L}(\mu)|_{X_w}$. This gives the following:

$$H^0\big(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu)\big) \simeq H^0(G/B, \mathcal{M}_w), \tag{6}$$

where $\mathcal{M}_w$ is the locally free sheaf on G/B associated to the B-module M_w. Now, by Theorem (5.7), we get by the isomorphism (6):

$$H^0\big(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu)\big) \simeq \bigoplus_{\theta \in \Lambda^+} V(\theta)^* \otimes \big[V(\theta) \otimes M_w\big]^{\mathfrak{b}} \tag{7}$$

$$\simeq \bigoplus_{\theta \in \Lambda^+} V(\theta)^* \otimes \mathrm{Hom}_{\mathfrak{b}}(M_w^*, V(\theta)). \tag{8}$$

Now, the theorem follows from the isomorphism:

$$H^0(X_w, \mathcal{L}_w(\mu))^* \simeq V_w(\mu), \text{ for } \mu \in \Lambda^+,$$

which of course is a consequence of the Demazure character formula (cf., e.g., [K$_4$, Corollary 8.1.26]). □

We recall the following result due to Joseph [Jo, §3.5], which is a generalization of Harish-Chandra's theorem used in the proof of Theorem (3.7).

Theorem (5.9). *For any $w \in W$ and $\mu \in \Lambda^+$, the map $U(\mathfrak{n}) \to V_w(\mu)$, defined by $X \mapsto X.v_{w\mu}$, has kernel precisely equal to the left $U(\mathfrak{n})$-ideal $\sum_{\alpha \in R^+} U(\mathfrak{n}) X_\alpha^{k_\alpha + 1}$, where X_α is any nonzero root vector in the root space $\mathfrak{g}_\alpha$ and k_α is defined as follows:*

$$k_\alpha = k_\alpha^\mu(w) = 0, \ \textit{if} (w\mu)(\alpha^\vee) \geq 0 \tag{9}$$

$$= -(w\mu)(\alpha^\vee), \ \textit{otherwise.} \tag{10}$$

Corollary (5.10). *For any $w \in W$ and $\lambda, \mu \in \Lambda^+$, $\mathrm{Hom}_{\mathfrak{b}}(\mathbb{C}_\lambda \otimes V_w(\mu), V(\overline{\lambda + w\mu}))$ is one-dimensional (over $\mathbb{C}$).*

Proof. Since $V_w(\mu)$ is a $U(\mathfrak{n})$-cyclic module generated by the element $v_{w\mu}$ of weight $w\mu$, $\mathbb{C}_\lambda$ is of weight λ, and the $\lambda + w\mu$ weight space in $V(\overline{\lambda + w\mu})$ is one-dimensional, we clearly have

$$\dim \mathrm{Hom}_{\mathfrak{b}}(\mathbb{C}_\lambda \otimes V_w(\mu), V(\overline{\lambda + w\mu})) \leq 1.$$

By the above theorem, the map $v_\lambda \otimes v_{w\mu} \mapsto v_{\lambda + w\mu}$ clearly extends uniquely to a $\mathfrak{b}$-module map, where $v_{\lambda + w\mu}$ is some fixed nonzero vector of weight $\lambda + w\mu$ in $V(\overline{\lambda + w\mu})$. □

For any $\lambda, \mu \in \Lambda^+$, by the Borel-Weil theorem, there is a G-module (in fact a $G \times G$-module) isomorphism $\xi : \big(V(\lambda) \otimes V(\mu)\big)^* \simeq H^0(G/B \times G/B, \mathcal{L}(\lambda \boxtimes \mu))$. On composition with the canonical restriction map $H^0(G/B \times G/B, \mathcal{L}(\lambda \boxtimes \mu)) \to H^0(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu))$, we get a G-module map

$$\xi_w : \big(V(\lambda) \otimes V(\mu)\big)^* \to H^0(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu)).$$

Since $\widetilde{C}_w = G \underset{B}{\times} C_w$ sits as a (Zariski) open subset of $\widetilde{X}_w$ (where $C_w = C_w^B$ is the Bruhat cell as in Section 2), the following lemma is trivial to prove.

Lemma (5.11). $\ker \xi_w = \{f \in \big(V(\lambda) \otimes V(\mu)\big)^* : f_{|U(\mathfrak{g})(v_\lambda \otimes v_{w\mu})} = 0\}$, *where* $U(\mathfrak{g})(v_\lambda \otimes v_{w\mu})$ *denotes the* $U(\mathfrak{g})$*-span of the vector* $v_\lambda \otimes v_{w\mu}$ *in* $V(\lambda) \otimes V(\mu)$.

By Corollary (5.5) (applied to $w = w_o$ and $v = w$), the map ξ_w is surjective and hence dualizing the above lemma, we get the following crucial:

Proposition (5.12). *For any* $w \in W$ *and* $\lambda, \mu \in \Lambda^+$,

$$H^0(\widetilde{X}_w, \mathcal{L}_w(\lambda \boxtimes \mu))^* \simeq U(\mathfrak{g})(v_\lambda \otimes v_{w\mu}) \hookrightarrow V(\lambda) \otimes V(\mu).$$

Now combining Theorem (5.8) with Corollary (5.10) and Proposition (5.12), we get the following most important result of this section. In the nineteen sixties, Parthasarathy-Ranga Rao-Varadarajan (for short PRV) conjectured (unpublished) the 'In particular' part of the following theorem (and proved it for $w = w_o$; cf. Theorem 5.1). Then, Kostant (in the mid eighties) came up with a more precise form of their conjecture (known as the PRVK conjecture), which is the first part of the following theorem. It was proved by Kumar [K_1] (using only char. 0 methods) and was extended by Mathieu [M_1] to an arbitrary char. The proof given here follows that of Kumar. Subsequently, other proofs of the original PRV conjecture appeared. Lusztig's results on the intersection homology of generalized Schubert varieties associated to affine Kac-Moody groups give a proof of the PRV conjecture; Rajeswari [Ra] gave a proof for classical G using Standard Monomial Theory; Littelmann [L_2] gave a proof using his LS path models.

Theorem (5.13). *For any finite-dimensional semisimple Lie algebra* $\mathfrak{g}$, *any* $\lambda, \mu \in \Lambda^+$, *and* $w \in W$, *the irreducible* $\mathfrak{g}$*-module* $V(\overline{\lambda + w\mu})$ *(with extremal weight* $\lambda + w\mu$*) occurs with multiplicity exactly one inside the* $\mathfrak{g}$*-submodule* $U(\mathfrak{g})(v_\lambda \otimes v_{w\mu})$ *of* $V(\lambda) \otimes V(\mu)$.

In particular, the $\mathfrak{g}$*-module* $V(\overline{\lambda + w\mu})$ *occurs with multiplicity at least one in* $V(\lambda) \otimes V(\mu)$.

Remark (5.14). (a) As proved in [K_1, Proposition 2.13], $V(\overline{\lambda + w\mu})$ occurs 'for the first time' in $U(\mathfrak{g})(v_\lambda \otimes v_{w\mu})$ if λ and μ are both regular. Precisely, for $\lambda, \mu \in \Lambda^{++}$, the $\mathfrak{g}$-module $V(\overline{\lambda + w\mu})$ does not occur in $U(\mathfrak{g})(v_\lambda \otimes v_{v\mu})$, for any $v < w$.

(b) Following recent works of Dimitrov-Roth and Ressayre (cf. [DR$_1$], [DR$_2$], [R$_3$]), one obtains the following: Let $\lambda, \mu, \nu \in \Lambda^+$ be such that there exists $w \in W$ with $\nu = \overline{\lambda + w\mu}$. Then, the following are equivalent:

(i) For all $k \geq 1$, $V(k\nu)$ appears in $V(k\lambda) \otimes V(k\mu)$ with multiplicity 1.

(ii) there exist $w_1, w_2, w_3 \in W$ such that $\ell(w_3) = \ell(w_1) + \ell(w_2)$, $w_3 * \nu = w_1 * \lambda + w_2 * \mu$ and the canonical product map

$$H^{\ell(w_1)}\big(G/B, \mathcal{L}(w_1 * \lambda)\big) \otimes H^{\ell(w_2)}\big(G/B, \mathcal{L}(w_2 * \mu)\big) \longrightarrow H^{\ell(w_3)}\big(G/B, \mathcal{L}(w_3 * \nu)\big)$$

is nonzero.

The following is a refinement of Theorem (5.13) proved by Kumar [K$_2$, Theorem 1.2] (which was conjectured by D.N. Verma).

Theorem (5.15). *Fix $\lambda, \mu \in \Lambda^+$ and consider the map $\eta : W_\lambda \backslash W / W_\mu \to \Lambda^+$, defined by $\eta(W_\lambda v W_\mu) = \overline{\lambda + v\mu}$, for any $v \in W$, where W_λ is the stabilizer of λ in W. Then, for any $w \in W$, the irreducible $\mathfrak{g}$-module $V(\overline{\lambda + w\mu})$ occurs in $V(\lambda) \otimes V(\mu)$ with multiplicity at least equal to $\#\eta^{-1}(\eta(W_\lambda w W_\mu))$, where $\#$ denotes the order.*

In particular, the number of irreducible components of $V(\lambda) \otimes V(\mu)$ (counted with multiplicities) is at least as much as the order of the double coset space $W_\lambda \backslash W / W_\mu$. (Of course, $W_\lambda = W_\mu = \{e\}$, if we assume λ and μ to be both regular.)

Proof. Fix a $w \in W$ and let $\{W_\lambda w_1 W_\mu, \ldots, W_\lambda w_n W_\mu\}$ be the distinct double cosets such that $\eta(W_\lambda w_i W_\mu) = \overline{\lambda + w\mu}$, for all $1 \leq i \leq n$, and such that each w_i is of minimal length in its double coset. By [BrK, Remark 3.1.3], the restriction map

$$H^0\big(G/B \times G/B, \mathcal{L}(\lambda \boxtimes \mu)\big) \to H^0(\mathcal{Y}, \mathcal{L}(\lambda \boxtimes \mu)_{|\mathcal{Y}})$$

is surjective, where $\mathcal{Y} := \bigcup_{i=1}^n \widetilde{X}_{w_i}$ is the closed subvariety (equipped with the reduced subscheme structure) of $G/B \times G/B$. For $1 \leq j \leq n$, define $\mathcal{Y}_j = \bigcup_{i=1}^j \widetilde{X}_{w_i}$. Now, the theorem follows from the following proposition due to Kumar [K$_2$, Proposition 2.5] together with [BrK, Exercise 3.3.E.3]. (This proposition is obtained by considering the ideal sheaf of $\mathcal{Y}_j$ in $\mathcal{Y}_{j+1}$ and induction on j.) □

Proposition (5.16). *For any $1 \leq j \leq n$, the irreducible $\mathfrak{g}$-module $V(\overline{\lambda + w\mu})$ occurs in $H^0\big(\mathcal{Y}_j, \mathcal{L}(\lambda \boxtimes \mu)_{|\mathcal{Y}_j}\big)^*$ with multiplicity exactly equal to j.*

6. Determination of the Saturated Tensor Cone

This section is based on the work [BK$_1$] due to Belkale-Kumar. We follow the notation and assumptions from Secton 2; in particular, G is a semisimple connected complex algebraic group.

For any $\lambda, \mu, \nu \in \Lambda^+$,

$$\mathrm{Hom}_G\big(V(\nu), V(\lambda) \otimes V(\mu)\big) \simeq [V(\lambda) \otimes V(\mu) \otimes V(\nu^*)]^G,$$

and hence the *tensor product problem* of determining the components $V(\nu)$ in the tensor product $V(\lambda) \otimes V(\mu)$ can be restated (replacing ν by ν^*) in a more symmetrical form of determining when $[V(\lambda) \otimes V(\mu) \otimes V(\nu)]^G \neq 0$. We generalize this problem from $s = 3$ to any $s \geq 1$ and define the *tensor product semigroup:*

$$\bar{\Gamma}_s(G) := \{(\lambda_1, \ldots, \lambda_s) \in (\Lambda^+)^s : [\lambda_1, \ldots, \lambda_s]^G \neq 0\},$$

where $[\lambda_1, \ldots, \lambda_s]^G$ denotes the dimension of the space of G-invariants $[V(\lambda_1) \otimes \cdots \otimes V(\lambda_s)]^G$. By Lemma (3.9), it is indeed a semigroup. Some general results on $\bar{\Gamma}_3(G)$ are obtained in the paper [KM$_1$] by Kapovich-Millson. The determination of $\bar{\Gamma}_s(G)$ in general is very hard, so we look at the weaker 'saturated tensor product problem' and define the *saturated tensor product semigroup*:

$$\Gamma_s(G) := \{(\lambda_1, \ldots, \lambda_s) \in (\Lambda^+)^s : [N\lambda_1, \ldots, N\lambda_s]^G \neq 0 \text{ for some } N > 0\}.$$

Let $\Lambda^+_{\mathbb{R}} := \{\lambda \in \Lambda \otimes_{\mathbb{Z}} \mathbb{R} : \lambda(\alpha_i^\vee) \geq 0 \text{ for all the simple coroots } \alpha_i^\vee\}$. By virtue of the convexity result in symplectic geometry, there exists a (unique) convex polyhedral cone $\Gamma_s(G)_{\mathbb{R}} \subset (\Lambda^+_{\mathbb{R}})^s$ such that

$$\Gamma_s(G) = \Gamma_s(G)_{\mathbb{R}} \cap \Lambda^s.$$

The aim of this section is to find the inequalities describing the cone $\Gamma_s(G)_{\mathbb{R}}$ explicitly. Observe that the cone $\Gamma_s(G)_{\mathbb{R}}$ depends only upon the Lie algebra $\mathfrak{g}$ of G.

The following deformation of the cohomology product in $H^*(G/P)$ is due to Belkale-Kumar [BK$_1$, §6]. This deformed product is crucially used in the determination of $\Gamma_s(G)$.

Definition (6.1). Let P be any standard parabolic subgroup of G. Write the standard cup product in $H^*(G/P, \mathbb{Z})$ in the $\{[X_w^P]\}$ basis as follows:

$$\left[X_u^P\right] \cdot \left[X_v^P\right] = \sum_{w \in W^P} c_{u,v}^w \left[X_w^P\right]. \tag{11}$$

Introduce the indeterminates τ_i for each $\alpha_i \in \Delta \setminus \Delta(P)$ and define a deformed cup product $\odot$ as follows:

$$\left[X_u^P\right] \odot \left[X_v^P\right] = \sum_{w \in W^P} \left(\prod_{\alpha_i \in \Delta \setminus \Delta(P)} \tau_i^{(w^{-1}\rho - u^{-1}\rho - v^{-1}\rho - \rho)(x_i)} \right) c_{u,v}^w \left[X_w^P\right],$$

where ρ is the (usual) half sum of positive roots of $\mathfrak{g}$ and x_i is defined in Section 2.

By (subsequent) Corollary (6.17), whenever $c_{u,v}^w$ is nonzero, the exponent of τ_i in the above is a nonnegative integer. Moreover, it is easy to see that the product $\odot$ is associative and clearly commutative. This product should not be confused with the small quantum cohomology of G/P.

The cohomology algebra of G/P obtained by setting each $\tau_i = 0$ in $(H^*(G/P,\mathbb{Z}) \otimes \mathbb{Z}[\tau_i], \odot)$ is denoted by $(H^*(G/P,\mathbb{Z}), \odot_0)$. Thus, as a $\mathbb{Z}$-module, it is the same as the singular cohomology $H^*(G/P,\mathbb{Z})$ and under the product $\odot_0$ it is associative (and commutative). Moreover, it continues to satisfy the Poincaré duality (cf. [BK$_1$, Lemma 16(d)]).

The cohomology algebra $H^*(G/P)$ under the product $\odot_0$ is intimately connected with the Lie algebra cohomology of the nil-radical $\mathfrak{u}$ of the parabolic subalgebra $\mathfrak{p}$ (cf. [BK$_1$, Theorem 43]).

We recall the following lemma from [BK$_1$, Lemma 19].

Lemma (6.2). *Let P be a cominuscule maximal standard parabolic subgroup of G (i.e., the simple root $\alpha_P \in \Delta \setminus \Delta(P)$ appears with coefficient 1 in the highest root of R^+). Then, the product $\odot$ coincides with the cup product in $H^*(G/P)$.*

Given a standard maximal parabolic subgroup P, let ω_P denote the corresponding fundamental weight, i.e., $\omega_P(\alpha_i^\vee) = 1$, if $\alpha_i \in \Delta \setminus \Delta(P)$ and 0 otherwise.

The following theorem due to Belkale-Kumar [BK$_1$, Theorem 22] determines the semigroup $\Gamma_s(G)$ 'most efficiently'. For $G = \mathrm{SL}(n)$, every maximal parabolic subgroup P is cominuscule and hence, by the above lemma, the deformed product $\odot_0$ in $H^*(G/P)$ coincides with the standard cup product. In this case, the following theorem was obtained by Klyachko [Kly] with a refinement by Belkale [B$_1$]. If we replace the product $\odot_0$ in (b) of the following theorem by the standard cup product, then the equivalence of (a) and (b) for general G was proved by Kapovich-Leeb-Millson [KLM] following an analogous slightly weaker result proved by Berenstein-Sjamaar [BS]. It may be mentioned that replacing the product $\odot_0$ in (b) by the standard cup product, we get, in general, 'far more' inequalities for simple groups other than SL_n. For example, for G of type B_3 (or C_3), the standard cup product gives rise to 135 inequalities, whereas the product $\odot_0$ gives only 102 inequalities (cf. [KuLM]).

Theorem (6.3). *Let $(\lambda_1, \dots, \lambda_s) \in (\Lambda^+)^s$. Then, the following are equivalent:*

(a) $(\lambda_1, \dots, \lambda_s) \in \Gamma_s(G)$.

(b) For every standard maximal parabolic subgroup P in G and every choice of s-tuples $(w_1, \dots, w_s) \in (W^P)^s$ such that

$$[X^P_{w_1}] \odot_0 \cdots \odot_0 [X^P_{w_s}] = [X^P_e] \in \big(H^*(G/P,\mathbb{Z}), \odot_0\big),$$

the following inequality holds:

$$\sum_{j=1}^{s} \lambda_j(w_j x_{i_P}) \le 0, \qquad (I^P_{(w_1,\dots,w_s)})$$

where α_{i_P} is the (unique) simple root in $\Delta \setminus \Delta(P)$.

The following result is due to Ressayre [R$_1$]. In the case $G = \mathrm{SL}(n)$, it was earlier proved by Knutson-Tao-Woodward [KTW].

Theorem (6.4). *The set of inequalities provided by the (b)-part of Theorem (6.3) is an irredundant system of inequalities describing the cone $\Gamma_s(G)_{\mathbb{R}}$ inside $(\Lambda^+_{\mathbb{R}})^s$, i.e., the hyperplanes given by the equality in $I^P_{(w_1,\dots,w_s)}$ are precisely those facets of the cone $\Gamma_s(G)_{\mathbb{R}}$ which intersect the interior of $(\Lambda^+_{\mathbb{R}})^s$.*

As a preparation towards the proof of Theorem (6.3), we first recall the following transversality theorem due to Kleiman (cf. [BK$_1$, Proposition 3]).

Theorem (6.5). *Let a connected algebraic group G act transitively on a smooth variety X and let $X_1, \dots, X_s$ be irreducible locally closed subvarieties of X. Then, there exists a nonempty open subset $U \subseteq G^s$ such that for $(g_1, \dots, g_s) \in U$, the intersection $\bigcap_{j=1}^s g_j X_j$ is proper (possibly empty) and dense in $\bigcap_{j=1}^s g_j \bar{X}_j$.*

Moreover, if X_j are smooth varieties, we can find such a U with the additional property that for $(g_1, \dots, g_s) \in U$, $\bigcap_{j=1}^s g_j X_j$ is transverse at each point of intersection.

We need the shifted Bruhat cell:

$$\Phi^P_w := w^{-1} B w P \subset G/P.$$

Let $T^P = T(G/P)_{\dot{e}}$ be the tangent space of G/P at $\dot{e} \in G/P$. It carries a canonical action of P. For $w \in W^P$, define T^P_w to be the tangent space of Φ^P_w at $\dot{e}$. We shall abbreviate T^P and T^P_w by T and T_w respectively when the reference to P is clear. By (4), B_L stabilizes Φ^P_w keeping $\dot{e}$ fixed. Thus,

$$B_L T_w \subset T_w. \tag{12}$$

The following result follows easily from the above transversality theorem and [F$_1$, Proposition 7.1 and §12.2] by observing that $g\Phi^P_w$ passes through $\dot{e} \Leftrightarrow g\Phi^P_w = p\Phi^P_w$ for some $p \in P$.

Proposition (6.6). *Take any $(w_1, \dots, w_s) \in (W^P)^s$ such that*

$$\sum_{j=1}^s \operatorname{codim} \Phi^P_{w_j} \le \dim G/P. \tag{13}$$

Then, the following three conditions are equivalent:

(a) $[X^P_{w_1}] \cdot \ldots \cdot [X^P_{w_s}] \neq 0 \in H^*(G/P)$.

(b) For generic $(p_1, \dots, p_s) \in P^s$, the intersection $p_1\Phi^P_{w_1} \cap \cdots \cap p_s\Phi^P_{w_s}$ is transverse at $\dot{e}$.

(c) For generic $(p_1, \dots, p_s) \in P^s$,

$$\dim(p_1 T_{w_1} \cap \cdots \cap p_s T_{w_s}) = \dim G/P - \sum_{j=1}^{s} \operatorname{codim} \Phi^P_{w_j}.$$

The set of s-tuples in (b) as well as (c) is an open subset of P^s.

The definition of the deformed product $\odot_0$ was arrived at from the following crucial concept.

Definition (6.7). Let $w_1, \dots, w_s \in W^P$ be such that

$$\sum_{j=1}^{s} \operatorname{codim} \Phi^P_{w_j} = \dim G/P. \tag{14}$$

We then call the s-tuple $(w_1, \dots, w_s)$ *Levi-movable* for short *L-movable* if, for generic $(l_1, \dots, l_s) \in L^s$, the intersection $l_1 \Phi_{w_1} \cap \cdots \cap l_s \Phi_{w_s}$ is transverse at $\dot{e}$.

By Proposition (6.6), if $(w_1, \dots, w_s)$ is L-movable, then $[X^P_{w_1}] \cdot \ldots \cdot [X^P_{w_s}] = d[X^P_e]$ in $H^*(G/P)$, for some nonzero d.

A Review of Geometric Invariant Theory. We need to consider the Geometric Invariant Theory (GIT) in a nontraditional setting, where a *nonreductive* group acts on a *nonprojective* variety. First we recall the following definition due to Mumford.

Definition (6.8). Let S be any (not necessarily reductive) algebraic group acting on a (not necessarily projective) variety $\mathbb{X}$ and let $\mathbb{L}$ be an S-equivariant line bundle on $\mathbb{X}$. Let $O(S)$ be the set of all one parameter subgroups (for short OPS) in S. Take any $x \in \mathbb{X}$ and $\delta \in O(S)$ such that the limit $\lim_{t\to 0} \delta(t)x$ exists in $\mathbb{X}$ (i.e., the morphism $\delta_x : \mathbb{G}_m \to X$ given by $t \mapsto \delta(t)x$ extends to a morphism $\widetilde{\delta}_x : \mathbb{A}^1 \to X$). Then, following Mumford, define a number $\mu^{\mathbb{L}}(x, \delta)$ as follows: Let $x_o \in X$ be the point $\widetilde{\delta}_x(0)$. Since x_o is $\mathbb{G}_m$-invariant via δ, the fiber of $\mathbb{L}$ over x_o is a $\mathbb{G}_m$-module; in particular, is given by a character of $\mathbb{G}_m$. This integer is defined as $\mu^{\mathbb{L}}(x, \delta)$.

We record the following standard properties of $\mu^{\mathbb{L}}(x, \delta)$ (cf. [MFK, Chap. 2, §1]):

Proposition (6.9). *For any* $x \in \mathbb{X}$ *and* $\delta \in O(S)$ *such that* $\lim_{t\to 0} \delta(t)x$ *exists in* $\mathbb{X}$*, we have the following (for any S-equivariant line bundles* $\mathbb{L}, \mathbb{L}_1, \mathbb{L}_2$*):*

(a) $\mu^{\mathbb{L}_1 \otimes \mathbb{L}_2}(x, \delta) = \mu^{\mathbb{L}_1}(x, \delta) + \mu^{\mathbb{L}_2}(x, \delta)$.

(b) If there exists $\sigma \in H^0(\mathbb{X}, \mathbb{L})^S$ *such that* $\sigma(x) \neq 0$*, then* $\mu^{\mathbb{L}}(x, \delta) \geq 0$.

(c) If $\mu^{\mathbb{L}}(x, \delta) = 0$*, then any element of* $H^0(\mathbb{X}, \mathbb{L})^S$ *which does not vanish at* x *does not vanish at* $\lim_{t\to 0} \delta(t)x$ *as well.*

(d) For any S-variety $\mathbb{X}'$ together with an S-equivariant morphism $f : \mathbb{X}' \to \mathbb{X}$ and any $x' \in \mathbb{X}'$ such that $\lim_{t\to 0} \delta(t)x'$ exists in $\mathbb{X}'$, we have $\mu^{f^*\mathbb{L}}(x', \delta) = \mu^{\mathbb{L}}(f(x'), \delta)$.

(e) (Hilbert-Mumford criterion) Assume that $\mathbb{X}$ is projective, S is connected and reductive and $\mathbb{L}$ is ample. Then, $x \in \mathbb{X}$ is semistable (with respect to $\mathbb{L}$) if and only if $\mu^{\mathbb{L}}(x, \delta) \geq 0$, for all $\delta \in O(S)$.

For an OPS $\delta \in O(S)$, let $\dot{\delta} \in \mathfrak{s}$ be its derivative at 1. Also, define the *associated parabolic subgroup* $P(\delta)$ of S by

$$P(\delta) := \left\{ g \in S : \lim_{t\to 0} \delta(t) g \delta(t)^{-1} \text{ exists in } S \right\}.$$

Definition (6.10). (Maximally destabilizing one parameter subgroups) We recall the definition of Kempf's OPS attached to an unstable point, which is in some sense 'most destabilizing' OPS. Let $\mathbb{X}$ be a projective variety with the action of a connected reductive group S and let $\mathbb{L}$ be a S-linearized ample line bundle on $\mathbb{X}$. Introduce the set $M(S)$ of fractional OPS in S. This is the set consisting of the ordered pairs (δ, a), where $\delta \in O(S)$ and $a \in \mathbb{Z}_{>0}$, modulo the equivalence relation $(\delta, a) \simeq (\gamma, b)$ if $\delta^b = \gamma^a$. The equivalence class of (δ, a) is denoted by $[\delta, a]$. An OPS δ of S can be thought of as the element $[\delta, 1] \in M(S)$. The group S acts on $M(S)$ via conjugation: $g \cdot [\delta, a] = [g\delta g^{-1}, a]$. Choose a S-invariant norm $q : M(S) \to \mathbb{R}_+$. We can extend the definition of $\mu^{\mathbb{L}}(x, \delta)$ to any element $\hat{\delta} = [\delta, a] \in M(S)$ and $x \in \mathbb{X}$ by setting $\mu^{\mathbb{L}}(x, \hat{\delta}) = \frac{\mu^{\mathbb{L}}(x,\delta)}{a}$. We note the following elementary property: If $\hat{\delta} \in M(S)$ and $p \in P(\delta)$ then

$$\mu^{\mathbb{L}}(x, \hat{\delta}) = \mu^{\mathbb{L}}(x, p\hat{\delta}p^{-1}). \tag{15}$$

For any unstable (i.e., nonsemistable) point $x \in \mathbb{X}$, define

$$q^*(x) = \inf_{\hat{\delta} \in M(S)} \{ q(\hat{\delta}) \mid \mu^{\mathbb{L}}(x, \hat{\delta}) \leq -1 \},$$

and the *optimal class*

$$\Lambda(x) = \{ \hat{\delta} \in M(S) \mid \mu^{\mathbb{L}}(x, \hat{\delta}) \leq -1, q(\hat{\delta}) = q^*(x) \}.$$

Any $\hat{\delta} \in \Lambda(x)$ is called *Kempf's OPS associated to x.*

By a theorem of Kempf (cf. [Ki, Lemma 12.13]), $\Lambda(x)$ is nonempty and the parabolic $P(\hat{\delta}) := P(\delta)$ (for $\hat{\delta} = [\delta, a]$) does not depend upon the choice of $\hat{\delta} \in \Lambda(x)$. The parabolic $P(\hat{\delta})$ for $\hat{\delta} \in \Lambda(x)$ will be denoted by $P(x)$ and called the *Kempf's parabolic associated to the unstable point x.*

We recall the following theorem due to Ramanan-Ramanathan [RR, Proposition 1.9].

Theorem (6.11). *For any unstable point $x \in \mathbb{X}$ and $\hat{\delta} = [\delta, a] \in \Lambda(x)$, let*

$$x_o = \lim_{t \to 0} \delta(t) \cdot x \in \mathbb{X}.$$

Then, x_o is unstable and $\hat{\delta} \in \Lambda(x_o)$.

Now, we return to the setting of Section 2. Let P be any standard parabolic subgroup of G acting on P/B_L via the left multiplication. We call $\delta \in O(P)$ *P-admissible* if, for all $x \in P/B_L$, $\lim_{t\to 0} \delta(t) \cdot x$ exists in P/B_L.

Observe that, B_L being the semidirect product of its commutator $[B_L, B_L]$ and T, any $\lambda \in \Lambda$ extends uniquely to a character of B_L. Thus, for any $\lambda \in \Lambda$, we have a P-equivariant line bundle $\mathcal{L}_P(\lambda)$ on P/B_L associated to the principal B_L-bundle $P \to P/B_L$ via the one-dimensional B_L-module λ^{-1}. The following lemma is easy to establish (cf. [BK$_1$, Lemma 14]). It is a generalization of the corresponding result in [BS, Section 4.2].

Lemma (6.12). *Let $\delta \in O(T)$ be such that $\dot{\delta} \in \mathfrak{t}_+ := \{x \in \mathfrak{t} : \alpha_i(x) \in \mathbb{R}_+ \, \forall$ the simple roots $\alpha_i\}$. Then, δ is P-admissible and, moreover, for any $\lambda \in \Lambda$ and $x = ulB_L \in P/B_L$ (for $u \in U, l \in L$), we have the following formula:*

$$\mu^{\mathcal{L}_P(\lambda)}(x, \delta) = -\lambda(w\dot{\delta}),$$

where $w \in W_P$ is the unique element such that $l^{-1} \in B_L w B_L$.

Definition (6.13). Let $w \in W^P$. Since T_w is a B_L-module (by (12)), we have the P-equivariant vector bundle $\mathcal{T}_w := P \underset{B_L}{\times} T_w$ on P/B_L. In particular, we have the P-equivariant vector bundle $\mathcal{T} := P \underset{B_L}{\times} T$ and $\mathcal{T}_w$ is canonically a P-equivariant subbundle of $\mathcal{T}$. Take the top exterior powers $\det(\mathcal{T}/\mathcal{T}_w)$ and $\det(\mathcal{T}_w)$, which are P-equivariant line bundles on P/B_L. Observe that, since T is a P-module, the P-equivariant vector bundle $\mathcal{T}$ is P-equivariantly isomorphic with the product bundle $P/B_L \times T$ under the map $\xi : P/B_L \times T \to \mathcal{T}$ taking $(pB_L, v) \mapsto [p, p^{-1}v]$, for $p \in P$ and $v \in T$; where P acts on $P/B_L \times T$ diagonally. We will often identify $\mathcal{T}$ with the product bundle $P/B_L \times T$ under ξ.

For $w \in W^P$, define the character $\chi_w \in \Lambda$ by

$$\chi_w = \sum_{\beta \in (R^+ \setminus R_{\mathfrak{l}}^+) \cap w^{-1}R^+} \beta \, .$$

Then, from [K$_4$, 1.3.22.3] and (4),

$$\chi_w = \rho - 2\rho^L + w^{-1}\rho, \tag{16}$$

where ρ (resp. ρ^L) is half the sum of roots in R^+ (resp. in $R_{\mathfrak{l}}^+$).

The following lemma is easy to establish.

Lemma (6.14). *For $w \in W^P$, as P-equivariant line bundles on P/B_L, we have:* $\det(\mathcal{T}/\mathcal{T}_w) = \mathcal{L}_P(\chi_w)$.

Let $\mathcal{T}_s$ be the P-equivariant product bundle $(P/B_L)^s \times T \to (P/B_L)^s$ under the diagonal action of P on $(P/B_L)^s \times T$. Then, $\mathcal{T}_s$ is canonically P-equivariantly isomorphic with the pull-back bundle $\pi_j^*(\mathcal{T})$, for any $1 \leq j \leq s$, where $\pi_j : (P/B_L)^s \to P/B_L$ is the projection onto the j-th factor. For any $w_1, \ldots, w_s \in W^P$, we have a P-equivariant map of vector bundles on $(P/B_L)^s$:

$$\Theta = \Theta_{(w_1,\ldots,w_s)} : \mathcal{T}_s \to \oplus_{j=1}^s \pi_j^*(\mathcal{T}/\mathcal{T}_{w_j}) \tag{17}$$

obtained as the direct sum of the projections $\mathcal{T}_s \to \pi_j^*(\mathcal{T}/\mathcal{T}_{w_j})$ under the identification $\mathcal{T}_s \simeq \pi_j^*(\mathcal{T})$. Now, assume that $w_1, \ldots, w_s \in W^P$ satisfies the condition (14). In this case, we have the same rank bundles on the two sides of the map (17). Let θ be the bundle map obtained from Θ by taking the top exterior power:

$$\theta = \det(\Theta) : \det\big(\mathcal{T}_s\big) \to \det\big(\mathcal{T}/\mathcal{T}_{w_1}\big) \boxtimes \cdots \boxtimes \det\big(\mathcal{T}/\mathcal{T}_{w_s}\big). \tag{18}$$

Clearly, θ is P-equivariant and hence one can view θ as a P-invariant element in

$$H^0\Bigg((P/B_L)^s, \det(\mathcal{T}_s)^* \otimes \Big(\det\big(\mathcal{T}/\mathcal{T}_{w_1}\big) \boxtimes \cdots \boxtimes \det\big(\mathcal{T}/\mathcal{T}_{w_s}\big)\Big)\Bigg).$$

$$= H^0\big((P/B_L)^s, \mathcal{L}_P(\chi_{w_1} - \chi_1) \boxtimes \mathcal{L}_P(\chi_{w_2}) \boxtimes \cdots \boxtimes \mathcal{L}_P(\chi_{w_s})\big). \tag{19}$$

The following lemma follows easily from Proposition (6.6).

Lemma (6.15). *Let $(w_1, \ldots, w_s)$ be an s-tuple of elements of W^P satisfying the condition (14). Then, we have the following:*

1. *The section θ is nonzero if and only if $[X_{w_1}^P] \cdot \ldots \cdot [X_{w_s}^P] \neq 0 \in H^*(G/P)$.*
2. *The s-tuple $(w_1, \ldots, w_s)$ is L-movable if and only if the section θ restricted to $(L/B_L)^s$ is not identically 0.*

Proposition (6.16). *Assume that $(w_1, \ldots, w_s) \in (W^P)^s$ satisfies condition (14). Then, the following are equivalent.*

(a) $(w_1, \ldots, w_s)$ is L-movable.

(b) $[X_{w_1}^P] \cdot \ldots \cdot [X_{w_s}^P] = d[X_e^P]$ in $H^*(G/P)$, for some nonzero d, and for each $\alpha_i \in \Delta \setminus \Delta(P)$, we have

$$\left(\left(\sum_{j=1}^s \chi_{w_j}\right) - \chi_1\right)(x_i) = 0.$$

Proof. (a)⇒(b): Let $(w_1, \dots, w_s) \in (W^P)^s$ be L-movable. Consider the restriction $\hat{\theta}$ of the P-invariant section θ to $(L/B_L)^s$. Then, $\hat{\theta}$ is non-vanishing by the above lemma. But, for

$$H^0\left((L/B_L)^s, (\mathcal{L}_P(\chi_{w_1} - \chi_1) \boxtimes \mathcal{L}_P(\chi_{w_2}) \boxtimes \cdots \boxtimes \mathcal{L}_P(\chi_{w_s}))_{|(L/B_L)^s}\right)^L$$

to be nonzero, the center of L should act trivially (under the diagonal action) on $\mathcal{L}_P(\chi_{w_1} - \chi_1) \boxtimes \mathcal{L}_P(\chi_{w_2}) \boxtimes \cdots \boxtimes \mathcal{L}_P(\chi_{w_s})$ restricted to $(L/B_L)^s$. This gives $\sum_{j=1}^s \chi_{w_j}(h) = \chi_1(h)$, for all h in the Lie algebra $\mathfrak{z}_L$ of the center of L; in particular, for $h = x_i$ with $\alpha_i \in \Delta \setminus \Delta(P)$.

(b)⇒(a): By the above lemma, $\theta(\bar{p}_1, \dots, \bar{p}_s) \neq 0$, for some $\bar{p}_j \in P/B_L$. Consider the central OPS of L: $\delta(t) := \prod_{\alpha_i \in \Delta \setminus \Delta(P)} t^{x_i}$. For any $x = ulB_L \in P/B_L$, with $u \in U$ and $l \in L$,

$$\lim_{t\to 0} \delta(t)x = \lim_{t\to 0} \delta(t)u\delta(t)^{-1}(\delta(t)l)B_L.$$

But, since $\beta(\dot{\delta}) > 0$, for all $\beta \in R^+ \setminus R_{\mathfrak{l}}^+$, we get $\lim_{t\to 0} \delta(t)u\delta(t)^{-1} = 1$. Moreover, since $\delta(t)$ is central in $L, \delta(t)lB_L$ equals lB_L. Thus, $\lim_{t\to 0} \delta(t)x$ exists and lies in L/B_L.

Now, let $\mathbb{L}$ be the P-equivariant line bundle $\mathcal{L}_P(\chi_{w_1} - \chi_1) \boxtimes \mathcal{L}_P(\chi_{w_2}) \boxtimes \cdots \boxtimes \mathcal{L}_P(\chi_{w_s})$ on $\mathbb{X} := (P/B_L)^s$, and $\bar{p} := (\bar{p}_1, \dots, \bar{p}_s) \in \mathbb{X}$. Then, by Lemma (6.12) (since δ is central in L), we get

$$\begin{aligned}\mu^{\mathbb{L}}(\bar{p}, \delta) &= -\sum_{\alpha_i \in \Delta \setminus \Delta(P)} \left(\left(\left(\sum_{j=1}^s \chi_{w_j}\right) - \chi_1\right)(x_i)\right) \\ &= 0, \text{ by assumption.}\end{aligned}$$

Therefore, using Proposition (6.9)(c) for $S = P$, θ does not vanish at $\lim_{t\to 0} \delta(t)\bar{p}$. But, from the above, this limit exists as an element of $(L/B_L)^s$. Hence, $(w_1, \dots, w_s)$ is L-movable by Lemma (6.15). □

Corollary (6.17). *For any $u, v, w \in W^P$ such that $c_{u,v}^w \neq 0$ (cf. equation (11)), we have*

$$(\chi_w - \chi_u - \chi_v)(x_i) \geq 0, \quad \textit{for each } \alpha_i \in \Delta \setminus \Delta(P). \tag{20}$$

Proof. By the assumption of the corollary, $[X_u^P] \cdot [X_v^P] \cdot [X^P_{w_o w w_o^P}] = d[X_e^P]$, for some nonzero d (in fact $d = c_{u,v}^w$), where $w_\circ^P$ is the longest element of W_P. Thus, by taking $(w_1, w_2, w_3) = (u, v, w_o w w_o^P)$ in Lemma (6.15), the section θ is nonzero. Now, apply Proposition (6.9)(b) for the OPS $\delta(t) = t^{x_i}$ and Lemma (6.12) (together with the identity (16)) to get the corollary. □

Proof of Theorem (6.3): Let $\mathbb{L}$ denote the G-linearized line bundle $\mathcal{L}(\lambda_1) \boxtimes \cdots \boxtimes \mathcal{L}(\lambda_s)$ on $(G/B)^s$ and let $P_1, \dots, P_s$ be the standard parabolic subgroups

such that $\mathbb{L}$ descends as an ample line bundle (still denoted by) $\mathbb{L}$ on $\mathbb{X} := G/P_1 \times \cdots \times G/P_s$. We call a point $x \in (G/B)^s$ *semistable* (with respect to, not necessarily ample, $\mathbb{L}$) if its image in $\mathbb{X}$ under the canonical map $\pi : (G/B)^s \to \mathbb{X}$ is semistable. Since the map π induces an isomorphism of G-modules:

$$H^0(\mathbb{X}, \mathbb{L}^N) \simeq H^0((G/B)^s, \mathbb{L}^N), \forall N > 0, \tag{21}$$

the condition (a) of Theorem (6.3) is equivalent to the following condition:

(c) The set of semistable points of $(G/B)^s$ with respect to $\mathbb{L}$ is nonempty.

Proof of the implication $(c) \Rightarrow (b)$ of Theorem (6.3): Let $x = (\bar{g}_1, \ldots, \bar{g}_s) \in (G/B)^s$ be a semistable point, where $\bar{g}_j = g_j B$. Since the set of semistable points is clearly open, we can choose a generic enough x such that the intersection $\cap\, g_j B w_j P$ itself is nonempty (cf. Theorem (6.5)). (By assumption, $\cap\, \overline{g_j B w_j P}$ is nonempty for any g_j.) Pick $f \in \cap\, g_j B w_j P$. Translating x by f^{-1}, we assume that $f = 1$. Consider the central OPS $\delta = t^{x_{i_P}}$ in L. Thus, applying Lemma (6.12) for $P = G$, the required inequality $I^P_{(w_1,\ldots,w_s)}$ is the same as $\mu^{\mathbb{L}}(x, \delta) \geq 0$; but this follows from Proposition (6.9), since x is semistable by assumption.

To prove the implication $(b) \Rightarrow (a)$ in Theorem (6.3), we need to recall the following result due to Kapovich-Leeb-Millson [KLM]. (For a self-contained algebro-geometric proof of this result, see [BK$_1$, §7.4].) Suppose that $x = (\bar{g}_1, \ldots, \bar{g}_s) \in (G/B)^s$ is an unstable point and $P(x)$ the Kempf's parabolic associated to $\pi(x)$. Let $\hat{\delta} = [\delta, a]$ be a Kempf's OPS associated to $\pi(x)$. Express $\delta(t) = f\gamma(t)f^{-1}$, where $\dot{\gamma} \in \mathfrak{t}_+$. Then, $P(\gamma)$ is a standard parabolic. Let P be a maximal parabolic containing $P(\gamma)$. Define $w_j \in W/W_{P(\gamma)}$ by $fP(\gamma) \in g_j B w_j P(\gamma)$ for $j = 1, \ldots, s$.

Theorem (6.18). *(i) The intersection $\bigcap_{j=1}^s g_j B w_j P \subset G/P$ is the singleton $\{fP\}$.*

(ii) For the simple root $\alpha_{i_P} \in \Delta \setminus \Delta(P)$, $\sum_{j=1}^s \lambda_j(w_j x_{i_P}) > 0$.

Now, we come to the proof of the implication $(b) \Rightarrow (a)$ in Theorem (6.3). Assume, if possible, that (a) (equivalently (c) as above) is false, i.e., the set of semistable points of $(G/B)^s$ is empty. Thus, any point $x = (\bar{g}_1, \ldots, \bar{g}_s) \in (G/B)^s$ is unstable. Choose a generic x so that for each standard parabolic $\widetilde{P}$ in G and any $(z_1, \ldots, z_s) \in W^s$, the intersection $g_1 B z_1 \widetilde{P} \cap \cdots \cap g_s B z_s \widetilde{P}$ is transverse (possibly empty) and dense in $g_1 \overline{B z_1 \widetilde{P}} \cap \cdots \cap g_s \overline{B z_s \widetilde{P}}$. Let $\hat{\delta} = [\delta, a], P, \gamma, f, w_j$ be as above associated to x. It follows from Theorem (6.18) that $\bigcap_{j=1}^s g_j B w_j P \subset G/P$ is the single point fP and, since x is generic, we get

$$[X^P_{w_1}] \cdot \ldots \cdot [X^P_{w_s}] = [X^P_e] \in H^*(G/P, \mathbb{Z}). \tag{22}$$

We now claim that the s-tuple $(w_1, \ldots, w_s) \in (W/W_P)^s$ is L-movable.

Write $g_j = f p_j w_j^{-1} b_j$, for some $p_j \in P(\gamma)$ and $b_j \in B$. Hence,

$$\delta(t)\bar{g}_j = f\gamma(t)p_j w_j^{-1} B = f\gamma(t)p_j\gamma^{-1}(t)w_j^{-1}B \in G/B.$$

Define, $l_j = \lim_{t\to 0} \gamma(t)p_j\gamma^{-1}(t)$. Then, $l_j \in L(\gamma)$, where $L(\gamma)$ is the Levi subgroup of $P(\gamma)$ containing T. Therefore,

$$\lim_{t\to 0} \delta(t)x = (fl_1w_1^{-1}B, \dots, fl_sw_s^{-1}B).$$

By Theorem (6.11), $\hat{\delta} \in \Lambda(\lim_{t\to 0}\delta(t)x)$. We further note that $fP(\gamma) \in \cap_j(fl_jw_j^{-1})Bw_jP(\gamma)$.

Applying Theorem (6.18) to the unstable point $x_o = \lim_{t\to 0}\delta(t)x$ yields: fP is the only point in the intersection $\bigcap_{j=1}^s fl_jw_j^{-1}Bw_jP$, i.e., translating by f, we get: $\dot{e} = eP$ is the only point in the intersection $\Omega := \bigcap_{j=1}^s l_jw_j^{-1}Bw_jP$. Thus, $\dim \Omega = 0$. By (22), the expected dimension of Ω is 0 as well. If this intersection Ω were not transverse at $\dot{e}$, then by [F_1, Remark 8.2], the local multiplicity at $\dot{e}$ would be > 1, each $w_j^{-1}Bw_jP$ being smooth. Further, G/P being a homogeneous space, any other component of the intersection $\bigcap \overline{l_jw_j^{-1}Bw_jP}$ contributes nonnegatively to the intersection product $[X^P_{w_1}] \cdot \ldots \cdot [X^P_{w_s}]$ (cf. [F_1, §12.2]). Thus, from (22), we get that the intersection $\bigcap l_jw_j^{-1}Bw_jP$ is transverse at $\dot{e} \in G/P$, proving that $(w_1, \dots, w_s)$ is L-movable. Thus, by Proposition (6.16) and the identities (16), (22), we get $[X^P_{w_1}] \odot_0 \dots \odot_0 [X^P_{w_s}] = [X^P_e]$. Now, part (ii) of Theorem (6.18) contradicts the inequality $I^P_{(w_1,\dots,w_s)}$. Thus, the set of semistable points of $(G/B)^s$ is nonempty, proving condition (a) of Theorem (6.3). □

Remark (6.19). (1) The cone $\Gamma_s(G)_{\mathbb{R}}$ coincides with the *eigencone* under the identification of $\mathfrak{t}_+$ with $\Lambda^+_{\mathbb{R}}$ induced from the Killing form (cf. [Sj, Theorem 7.6]). The eigencone for $G = \mathrm{SL}(n)$ has extensively been studied since the initial work of H. Weyl in 1912. For a detailed survey on the subject, we refer to Fulton's article [F_2].

(2) The cone $\Gamma_3(G)_{\mathbb{R}}$ is quite explicitly determined for any semisimple G of rank 2 in [KLM, §7], any simple G of rank 3 in [KuLM] and for $G = \mathrm{Spin}(8)$ in [KKM]. It has $50, 102, 102, 306$ facets for G of type A_3, B_3, C_3, D_4 respectively.

(3) The 'explicit' determination of $\Gamma_s(G)$ via Theorem (6.3) hinges upon understanding the product $\odot_0$ in $H^*(G/P)$ in the Schubert basis, for all the maximal parabolic subgroups P. Clearly, the product $\odot_0$ is easier to understand than the usual cup product (which is the subject matter of *Schubert Calculus*) since, in general, 'many more' terms in the product $\odot_0$ in the Schubert basis drop out. For the lack of space, we do not recall various results about the product $\odot_0$, instead we refer to the papers [BK_1, §§9,10], [BK_2, §§8,9], [KKM, §4], [PS], [Ri_1], [Ri_2], [ReR], [R_3].

7. Special Isogenies and Tensor Product Multiplicities

This section is based on the work [KS] due to Kumar-Stembridge. It exploits certain 'exceptional' isogenies between semisimple algebraic groups over alge-

braically closed fields of char. $p > 0$ to derive relations between tensor product multiplicities of $\mathrm{Spin}_{2\ell+1}$ and $\mathrm{Sp}_{2\ell}$ and also between two different sets of multiplicities of F_4 (and also that of G_2).

Let $G = G(k)$ and $G' = G'(k)$ be connected, semisimple algebraic groups over an algebraically closed field k of char. $p > 0$, and let $f : G \to G'$ be an isogeny (i.e., a surjective algebraic group homomorphism with a finite kernel). Fix a Borel subgroup B of G and $T \subset B$ a maximal torus, and let $B' = f(B)$ and $T' = f(T)$ be the corresponding groups in G'. Then, T' (resp. B') is a maximal torus (resp. a Borel subgroup) of G'.

The map f induces a homomorphism $f^* : \Lambda(T') \to \Lambda(T)$, which extends to an isomorphism $f^*_{\mathbb{R}} : \Lambda(T')_{\mathbb{R}} \xrightarrow{\sim} \Lambda(T)_{\mathbb{R}}$, where $\Lambda(T)_{\mathbb{R}} := \Lambda(T) \otimes_{\mathbb{Z}} \mathbb{R}$. Moreover, f^* takes $\Lambda(T')^+$ to $\Lambda(T)^+$.

Letting $R = R(G, T)$ denote the root system of G with respect to T and similarly $R' = R(G', T')$, we recall the following from [C, Exposé n° 18, Definition 1].

Definition (7.1). An isomorphism $\phi : \Lambda(T')_{\mathbb{R}} \to \Lambda(T)_{\mathbb{R}}$ is called *special* if $\phi(\Lambda(T')) \subset \Lambda(T)$, and there exist integers $d(\alpha) \geq 0$ such that $R' = \{p^{d(\alpha)}\phi^{-1}(\alpha) : \alpha \in R\}$.

For any isogeny f as above, the induced map $f^*_{\mathbb{R}}$ is a special isomorphism. Conversely, for any special isomorphism $\phi : \Lambda(T')_{\mathbb{R}} \to \Lambda(T)_{\mathbb{R}}$, there exists an isogeny $f : G \to G'$ with $f^*_{\mathbb{R}} = \phi$ (cf. [C, Exposé n° 23, §3, Théorème 1]).

In the following, an important result due to Donkin, asserting the existence of good filtrations for tensor products of the space of global sections of homogeneous line bundles, has been used. (It should be noted that Donkin proved this result for almost all the cases barring a few exceptions involving small primes [D]; the result was subsequently proved uniformly by Mathieu for all primes [M_2].) This allows replacing the following inequality (23) with a cohomological statement that is independent of the char. of the field (including the char. 0 case), thereby enabling us to deduce the inequality directly from the existence of an isogeny in char. p.

Theorem (7.2). *If $f : G \to G'$ is an isogeny of connected semisimple algebraic groups over an algebraically closed field k of char. $p > 0$, then for all $\lambda'_1, \ldots, \lambda'_n \in \Lambda(T')^+$,*

$$[\lambda'_1, \ldots, \lambda'_n]^{G'(\mathbb{C})} \leq [f^*(\lambda'_1), \ldots, f^*(\lambda'_n)]^{G(\mathbb{C})}, \tag{23}$$

where $G(\mathbb{C})$ is the connected semisimple complex algebraic group with the same root datum as that of $G(k)$ and similarly for $G'(\mathbb{C})$.

Proof. The map f clearly induces a surjective morphism (of varieties) $\bar{f} : X_n \to X'_n$, where $X_n := (G/B)^{\times n}$. Consider the dominant line bundle $\mathcal{L}(\lambda'_1) \boxtimes \cdots \boxtimes \mathcal{L}(\lambda'_n)$ on X'_n. Then, the pull-back line bundle on X_n is the homogeneous line bundle $\mathcal{L}(\lambda_1) \boxtimes \cdots \boxtimes \mathcal{L}(\lambda_n)$, where $\lambda_i := f^*(\lambda'_i)$. Thus, we get an injective map

$$\bar{f}^* : H^0\big(X'_n, \mathcal{L}(\lambda'_1) \boxtimes \cdots \boxtimes \mathcal{L}(\lambda'_n)\big) \hookrightarrow H^0\big(X_n, \mathcal{L}(\lambda_1) \boxtimes \cdots \boxtimes \mathcal{L}(\lambda_n)\big).$$

Since the map $\bar{f}$ is f-equivariant under the diagonal action of G on X_n and G' on X'_n, the injection $\bar{f}^*$ induces an injection (still denoted by)

$$\bar{f}^* : H^0\big(X'_n, \mathcal{L}(\lambda'_1) \boxtimes \cdots \boxtimes \mathcal{L}(\lambda'_n)\big)^{G'} \hookrightarrow H^0\big(X_n, \mathcal{L}(\lambda_1) \boxtimes \cdots \boxtimes \mathcal{L}(\lambda_n)\big)^{G}. \quad (24)$$

We have of course

$$H^0\big(X_n, \mathcal{L}(\lambda_1) \boxtimes \cdots \boxtimes \mathcal{L}(\lambda_n)\big) \cong H^0\big(G/B, \mathcal{L}(\lambda_1)\big) \otimes \cdots \otimes H^0\big(G/B, \mathcal{L}(\lambda_n)\big).$$

By [BrK, Corollary 4.2.14], the above module $M := H^0\big(G/B, \mathcal{L}(\lambda_1)\big) \otimes \cdots \otimes H^0\big(G/B, \mathcal{L}(\lambda_n)\big)$ admits a good filtration. Hence, by [BrK, Theorem 4.2.7, identity (4.2.1.3) and Proposition 4.2.3(c)], its T-character is

$$\operatorname{ch} M = \sum_{\lambda \in \Lambda(T)^+} \dim\big[H^0\big(G/B, \mathcal{L}(\lambda)\big) \otimes M\big]^G \cdot \operatorname{ch}(V_k(\lambda)),$$

where $V_k(\lambda) := H^0(G/B, \mathcal{L}(\lambda))^*$ is the Weyl module with highest weight λ. Recall that, by the Borel-Weil Theorem, $H^0\big(G(\mathbb{C})/B(\mathbb{C}), \mathcal{L}_{\mathbb{C}}(\lambda)\big) \simeq V(\lambda)^*$, where (as earlier) $V(\lambda)$ is the (complex) irreducible $G(\mathbb{C})$-module with highest weight λ and $\mathcal{L}_{\mathbb{C}}(\lambda)$ is the homogeneous line bundle on $G(\mathbb{C})/B(\mathbb{C})$ corresponding to the character λ^{-1} of $B(\mathbb{C})$. Moreover, as is well-known, $\operatorname{ch} V_k(\lambda) = \operatorname{ch} V(\lambda)$. (This follows from the vanishing of the cohomology $H^i(G/B, \mathcal{L}(\lambda))$ for all $i > 0$.)

But, clearly, $\operatorname{ch} M = \operatorname{ch}(V_k(\lambda_1)^*) \cdots \operatorname{ch}(V_k(\lambda_n)^*)$; in particular, it is independent of the char. of the field (including char. 0). Moreover, since $\{\operatorname{ch} V(\lambda)\}_{\lambda \in \Lambda(T)^+}$ are $\mathbb{Z}$-linearly independent as elements of the group ring of $\Lambda(T)$, we deduce that $\dim\big[H^0\big(G/B, \mathcal{L}(\lambda)\big) \otimes M\big]^G$ is independent of the char. of the base field for all $\lambda \in \Lambda(T)^+$. Taking $\lambda = 0$, we obtain that $\dim M^G$ is independent of the char. Observe next that (24) implies

$$\dim[M']^{G'} \leq \dim[M]^G, \quad (25)$$

where $M' := H^0\big(G'/B', \mathcal{L}(\lambda'_1)\big) \otimes \cdots \otimes H^0\big(G'/B', \mathcal{L}(\lambda'_n)\big)$.

Thus, (25) implies

$$\begin{aligned} \dim\big[V(\lambda'_1) \otimes \cdots \otimes V(\lambda'_n)\big]^{G'(\mathbb{C})} &= \dim\big[V(\lambda'_1)^* \otimes \cdots \otimes V(\lambda'_n)^*\big]^{G'(\mathbb{C})} \\ &\leq \dim\big[V(\lambda_1)^* \otimes \cdots \otimes V(\lambda_n)^*\big]^{G(\mathbb{C})} = \dim[V(\lambda_1) \otimes \cdots \otimes V(\lambda_n)]^{G(\mathbb{C})}. \end{aligned}$$

□

Definition (7.3). An isogeny $f : G \to G'$ for a simple G is called *special* if $d(\alpha) = 0$ for some $\alpha \in R(G, T)$, where $d(\alpha)$ is as in Definition (7.1); it is *central* if $d(\alpha) = 0$ for all $\alpha \in R(G, T)$. A complete list of special non-central isogenies may be found in [BT, §3.3]. In the following, we list the resulting tensor product inequalities implied by Theorem (7.2).

Let G be the simply-connected group of type B_ℓ (i.e., $G = \mathrm{Spin}_{2\ell+1}$), and G' the simply-connected group of type C_ℓ (i.e., $G' = \mathrm{Sp}_{2\ell}$). Following the notation from the appendices of [Bo], we identify $\Lambda(T) = \{\sum_{i=1}^{\ell} a_i\varepsilon_i : a_i \pm a_j \in \mathbb{Z}\,\forall i,j\}$ and $\Lambda(T') = \bigoplus_{i=1}^{\ell} \mathbb{Z}\varepsilon_i$. This provides a canonical inclusion $\Lambda(T') \hookrightarrow \Lambda(T)$, $\varepsilon_i \mapsto \varepsilon_i$, which takes $\Lambda(T')^+ \hookrightarrow \Lambda(T)^+$. Moreover, under this identification, the image of $\Lambda(T')$ (resp. $\Lambda(T')^+$) is precisely equal to $\Lambda(\bar{T})$ (resp. $\Lambda(\bar{T})^+$), where $\bar{T}$ is the maximal torus in $\mathrm{SO}_{2\ell+1}$.

Theorem (7.2) specializes as follows.

Corollary (7.4). *(a) If $\lambda_1, \ldots, \lambda_n$ are dominant weights for $\mathrm{Sp}_{2\ell}$ ($\ell \geq 2$), then*

$$[\lambda_1, \ldots, \lambda_n]^{\mathrm{Sp}_{2\ell}(\mathbb{C})} \leq [\lambda_1, \ldots, \lambda_n]^{\mathrm{SO}_{2\ell+1}(\mathbb{C})}.$$

(b) If $\lambda_1, \ldots, \lambda_n$ are dominant weights for $\mathrm{Spin}_{2\ell+1}$ ($\ell \geq 2$), then

$$[\lambda_1, \ldots, \lambda_n]^{\mathrm{Spin}_{2\ell+1}(\mathbb{C})} \leq [2\lambda_1, \ldots, 2\lambda_n]^{\mathrm{Sp}_{2\ell}(\mathbb{C})}.$$

(c) If $\lambda_1, \ldots, \lambda_n$ are dominant weights for F_4, then

$$[\lambda_1, \ldots, \lambda_n]^{F_4(\mathbb{C})} \leq [\phi(\lambda_1), \ldots, \phi(\lambda_n)]^{F_4(\mathbb{C})},$$

where $\phi(a\omega_1+b\omega_2+c\omega_3+d\omega_4) := d\omega_1+c\omega_2+2b\omega_3+2a\omega_4$ (ω_i being fundamental weights).

(d) If $\lambda_1, \ldots, \lambda_n$ are dominant weights for G_2, then

$$[\lambda_1, \ldots, \lambda_n]^{G_2(\mathbb{C})} \leq [\phi(\lambda_1), \ldots, \phi(\lambda_n)]^{G_2(\mathbb{C})},$$

where $\phi(a\omega_1 + b\omega_2) := 3b\omega_1 + a\omega_2$.

Proof. (a) The identity map is a special isomorphism $\Lambda(T')_{\mathbb{R}} \to \Lambda(T)_{\mathbb{R}}$ giving rise to an isogeny $f : \mathrm{SO}_{2\ell+1}(k) \to \mathrm{Sp}_{2\ell}(k)$, where char. $k = 2$.

(b) In this case, the map $\mu \mapsto 2\mu$ defines a special isomorphism $\Lambda(T')_{\mathbb{R}} \to \Lambda(T)_{\mathbb{R}}$ inducing an isogeny $f : \mathrm{Sp}_{2\ell}(k) \to \mathrm{Spin}_{2\ell+1}(k)$, where char. $k = 2$.

(c) In this case, the simple roots generate $\Lambda(T)$. Numbering them $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ as in [Bo], we have that α_1 and α_2 are long and α_3 and α_4 are short. Then, there is a special isomorphism $\phi : \Lambda(T)_{\mathbb{R}} \to \Lambda(T)_{\mathbb{R}}$ such that

$$\phi(\alpha_1) = 2\alpha_4,\ \phi(\alpha_2) = 2\alpha_3,\ \phi(\alpha_3) = \alpha_2,\ \phi(\alpha_4) = \alpha_1.$$

Let $G = G'$ be of type F_4 and char. $k = 2$ and apply Theorem (7.2).

(d) Letting α_1 and α_2 denote the simple roots, with α_1 short and α_2 long, there is a special isomorphism $\phi : \Lambda(T)_{\mathbb{R}} \to \Lambda(T)_{\mathbb{R}}$ such that $\phi(\alpha_1) = \alpha_2$, $\phi(\alpha_2) = 3\alpha_1$. Let $G = G'$ be of type G_2 and char. $k = 3$ and apply Theorem (7.2). □

As an immediate corollary of the (a) and the (b) parts above, we have the following:

Corollary (7.5). *For any $s \geq 1$ and any $\ell \geq 2$, the saturated tensor semigroup $\Gamma_s(\mathrm{Sp}_{2\ell}(\mathbb{C})) = \Gamma_s(\mathrm{SO}_{2\ell+1}(\mathbb{C}))$ under the identification of their $\Lambda(T)^+$ as above.*

Remark (7.6). (a) Any nonspecial isogenies or central isogenies do not yield any new inequalities.

(b) There is another combinatorial proof of Theorem (7.2) based on Littelmann's Path Model for tensor product multiplicity. More specifically, Kumar-Stembridge [KS] use a variant of the Path Model (see [St_2]) in which the objects are chains in the Bruhat ordering of various Weyl group orbits, and the inequality is obtained by comparing chains related by integer renormalizations.

8. Saturation Problem

We continue to follow the notation and assumptions from Secton 2; in particular, G is a semisimple connected complex algebraic group. In Section 6, we defined the tensor product semigroup $\bar{\Gamma}_s(G)$ as well as the saturated tensor product semigroup $\Gamma_s(G)$ (for any integer $s \geq 1$) and determined $\Gamma_s(G)$ by describing its facets. The *saturation problem* aims at connecting these two semigroups.

We begin with the following definition. We take $s = 3$ as this is the most relevant case to the tensor product decomposition.

Definition (8.1). An integer $d \geq 1$ is called a *saturation factor* for G, if for any $(\lambda, \mu, \nu) \in \Gamma_3(G)$ such that $\lambda + \mu + \nu \in Q$, $(d\lambda, d\mu, d\nu) \in \bar{\Gamma}_3(G)$, where Q is the root lattice of G. Of course, if d is a saturation factor then so is its any multiple. If $d = 1$ is a saturation factor for G, we say that the *saturation property holds for G.*

The *saturation theorem* of Knutson-Tao [KT], proved by using their 'honeycomb model' asserts the following. Other proofs of their result are given by Derksen-Weyman [DK], Belkale [B_2] and Kapovich-Millson [KM_2] (cf. Theorem (8.3) below).

Theorem (8.2). *The saturation property holds for $G = \mathrm{SL}(n)$.*

The following general result (though not optimal) on saturation factor is obtained by Kapovich-Millson [KM_2] by using the geometry of geodesics in Euclidean buildings and Littelmann's path model. A weaker form of the following theorem was conjectured by Kumar in a private communication to J. Millson (also see [KT, Conjecture]).

Theorem (8.3). *For any connected simple G, $d = k_{\mathfrak{g}}^2$ is a saturated factor, where $k_{\mathfrak{g}}$ is the least common multiple of the coefficients of the highest root θ of the Lie algebra $\mathfrak{g}$ of G written in terms of the simple roots $\{\alpha_1, \ldots, \alpha_\ell\}$.*

Observe that the value of $k_{\mathfrak{g}}$ is 1 for $\mathfrak{g}$ of type $A_\ell(\ell \geq 1)$; it is 2 for $\mathfrak{g}$ of type $B_\ell(\ell \geq 2), C_\ell(\ell \geq 3), D_\ell(\ell \geq 4)$; and it is 6, 12, 60, 12, 6 for $\mathfrak{g}$ of type E_6, E_7, E_8, F_4, G_2 respectively.

Kapovich-Millson determined $\bar{\Gamma}_3(G)$ explicitly for $G = \mathrm{Sp}(4)$ and G_2 (cf. [KM_1, Theorems 5.3, 6.1]). In particular, from their description, the following theorem follows easily.

Theorem (8.4). *The saturation property does not hold for either $G = \mathrm{Sp}(4)$ or G_2. Moreover, 2 is a saturation factor (and no odd integer d is a saturation factor) for $\mathrm{Sp}(4)$, whereas both of $2, 3$ are saturation factors for G_2 (and hence any integer $d > 1$ is a saturation factor for G_2).*

It was known earlier that the saturation property fails for G of type B_ℓ (cf. [E]).

Kapovich-Millson [KM_1] made the following very interesting conjecture:

Conjecture (8.5). *If G is simply-laced, then the saturation property holds for G.*

Apart from $G = \mathrm{SL}(n)$, the only other simply-connected, simple, simply-laced group G for which the above conjecture is known so far is $G = \mathrm{Spin}(8)$, proved by Kapovich-Kumar-Millson [KKM, Theorem 5.3] by explicit calculation using Theorem (6.3).

Theorem (8.6). *The above conjecture is true for $G = \mathrm{Spin}(8)$.*

Finally, we have the following improvement of Theorem (8.3) for the groups $\mathrm{SO}(2\ell + 1)$ and $\mathrm{Sp}(2\ell)$ due to Belkale-Kumar [BK_2, Theorems 25 and 26].

Theorem (8.7). *For the groups $\mathrm{SO}(2\ell+1)$ and $\mathrm{Sp}(2\ell)$, 2 is a saturation factor.*

The proof of the above theorem relies on the following theorem [BK_2, Theorem 23].

Theorem (8.8). *Let $(\lambda^1, \ldots, \lambda^s) \in \bar{\Gamma}_s(\mathrm{SL}(2\ell))$. Then, $(\lambda^1_C, \ldots, \lambda^s_C) \in \bar{\Gamma}_s(\mathrm{Sp}(2\ell))$, where λ^j_C is the restriction of λ^j to the maximal torus of $\mathrm{Sp}(2\ell)$.*

A similar result is true for $\mathrm{Sp}(2\ell)$ replaced by $\mathrm{SO}(2\ell + 1)$.

Belkale-Kumar [BK_2, Conjecture 29] conjectured the following generalization of Theorem (8.8). Let G be a simply-connected, semisimple complex algebraic group and let σ be a diagram automorphism of G with fixed subgroup $G^\sigma = K$.

Conjecture (8.9). *Let $(\lambda^1, \ldots, \lambda^s) \in \bar{\Gamma}_s(G)$. Then, $(\lambda^1_K, \ldots, \lambda^s_K) \in \bar{\Gamma}_s(K)$, where λ^j_K is the restriction of λ^j to the maximal torus of K.*

(Observe that, for any dominant character λ for G, λ_K is dominant for K with respect to the Borel subgroup $B^K := B^\sigma$ of K.)

We also mention the following 'rigidity' result (conjectured by Fulton) due to Knutson-Tao-Woodward [KTW] proved by combinatorial methods. There are now geometric proofs of the theorem by Belkale [B_3] and Ressayre [R_2].

Theorem (8.10). *Let $G = \mathrm{SL}(n)$ and let $\lambda, \mu, \nu \in \Lambda^+$. If $[V(\lambda)\otimes V(\mu)\otimes V(\nu)]^G$ is one-dimensional then so is $[V(N\lambda) \otimes V(N\mu) \otimes V(N\nu)]^G$, for any $N \geq 1$.*

The direct generalization of this theorem for other groups is, in general, false. But, a certain cohomological reinterpretation of the theorem remains true for any G (cf. a forthcoming paper by Belkale-Kumar-Ressayre).

9. Generalization of Littlewood-Richardson Formula

We recall the classical Littlewood-Richardson formula for $\mathrm{GL}(n)$ (cf., e.g., [Ma, Chap. 1, §9]). Let T be the standard maximal torus of $\mathrm{GL}(n)$ consisting of invertible diagonal matrices. Then, the irreducible polynomial representations of $\mathrm{GL}(n)$ (i.e., those irreducible representations whose matrix coefficients extend as a regular function on the whole of $M(n)$) are parametrized by the partitions $\lambda : (\lambda_1 \geq \cdots \geq \lambda_n \geq 0)\,(\lambda_i \in \mathbb{Z})$, where λ is viewed as an element of Λ^+ via the character: $\mathrm{diag}(t_1, \ldots, t_n) \mapsto t_1^{\lambda_1} \ldots t_n^{\lambda_n}$. Consider the decomposition (1) in Section 1 for the tensor product of irreducible polynomial representations of $\mathrm{GL}(n)$.

Theorem (9.1). *$m^\nu_{\lambda,\mu} \neq 0$ only if both of $\lambda, \mu \subset \nu$. In this case, $m^\nu_{\lambda,\mu}$ equals the number of tableaux T of shape $\nu - \lambda$ and weight μ such that the word $w(T) = (a_1, \ldots, a_N)$ associated to T (reading the symbols in T from right to left in successive rows starting with the top row) is a lattice permutation, i.e., for all $1 \leq i \leq m-1$, and $1 \leq r \leq N, \#\{j \leq r : a_j = i\} \geq \#\{j \leq r : a_j = i+1\}$, where the symbols in T lie in $\{1, \ldots, m\}$.*

Littelmann generalized the above thorem for all semisimple Lie algebras $\mathfrak{g}$ by using his LS path models as below. Let G be the simply-connected complex algebraic group with Lie algebra $\mathfrak{g}$.

Definition (9.2). Let Π be the set of all piecewise-linear, continuous paths $\gamma : [0, 1] \to \Lambda_\mathbb{R} := \Lambda \otimes_\mathbb{Z} \mathbb{R}$ with $\gamma(0) = 0$ and $\gamma(1) \in \Lambda$, modulo the equivalence relation $\gamma \equiv \gamma'$ if γ' is obtained from γ by a piecewise-linear, nondecreasing, continuous reparametrization. For any simple root α_i, there are two operators $e_{\alpha_i}, f_{\alpha_i} : \Pi \sqcup \{0\} \to \Pi \sqcup \{0\}$ defined in [L_2], [L_3]. Let Π^+ be the set of those paths $\gamma \in \Pi$ such that $\mathrm{Im}\ \gamma \subset \Lambda^+_\mathbb{R}$. For any $\gamma \in \Pi^+$, let $\mathcal{P}_\gamma$ be the smallest subset of Π containing γ such that $\mathcal{P}_\gamma \sqcup \{0\}$ is stable under the operators $\{e_{\alpha_i}, f_{\alpha_i}; 1 \leq i \leq \ell\}$.

The following theorem due to Littelmann [L_2], [L_3] generalizes Theorem (9.1).

Theorem (9.3). *For any $\lambda, \mu \in \Lambda^+$, take any path $\gamma_\lambda, \gamma_\mu \in \Pi^+$ such that $\gamma_\lambda(1) = \lambda$ and $\gamma_\mu(1) = \mu$. Then,*

$$V(\lambda) \otimes V(\mu) = \bigoplus_{\gamma} V(\lambda + \gamma(1)),$$

where γ runs over all the paths in $\mathcal{P}_{\gamma_\mu}$ such that the cancatenation $\gamma_\lambda * \gamma \in \Pi^+$.

By [L_2, §8] (also see [L_1]), the above theorem indeed generalizes Theorem (9.1).

We now come to the tensor product multiplicity formula due to Berenstein-Zelevinsky [BZ, Theorem 2.3].

Definition (9.4). Let V be a finite-dimensional representation of G and let $\lambda, \mu \in P(V)$ (the set of weights of V), and let $\mathbf{i} = (i_1, \ldots, i_r)$ be a sequence with $1 \le i_j \le \ell$. An $\mathbf{i}$-*trail* from λ to μ in V is a sequence of weights $\mathcal{T} = (\lambda_0 = \lambda, \lambda_1, \ldots, \lambda_r = \mu)$ in $P(V)$ such that

(1) for all $1 \le j \le r$, we have $\lambda_{j-1} - \lambda_j = c_j(\mathcal{T})\alpha_{i_j}$, for some $c_j = c_j(\mathcal{T}) \in \mathbb{Z}_+$, and

(2) $e_{i_1}^{c_1} \ldots e_{i_r}^{c_r} : V_\mu \to V_\lambda$ is a nonzero map, where e_{i_j} is a nonzero simple root vector as in Section 2 and V_μ is the weight space of V corresponding to the weight μ.

Fix a reduced word for the longest element $w_o = s_{i_1} \ldots s_{i_N}$ and let $\mathbf{i_o} = (i_1, \ldots, i_N)$.

Theorem (9.5). *For $\lambda, \mu, \nu \in \Lambda^+$, the tensor product multiplicity $m_{\lambda,\mu}^{\nu}$ equals the number of N-tuples $(d_1, \ldots, d_N)$ of nonnegative integers satisfying the following conditions:*

(a) $\sum_{j=1}^{N} d_j s_{i_1} \ldots s_{i_{j-1}} \alpha_{i_j} = \lambda + \mu - \nu$,

(b) $\sum_j c_j(\mathcal{T}) d_j \ge (s_i\lambda + \mu - \nu)(\omega_i^\vee)$, *for any* $1 \le i \le \ell$ *and any* $\mathbf{i_o}$*-trail* $\mathcal{T}$ *from* $s_i\omega_i^\vee$ *to* $w_o\omega_i^\vee$ *in* $V(\omega_i^\vee)$, *and*

(c) $\sum_j c_j(\mathcal{T}) d_j \ge (\lambda + s_i\mu - \nu)(\omega_i^\vee)$, *for any* $1 \le i \le \ell$ *and any* $\mathbf{i_o}$*-trail* $\mathcal{T}$ *from* $\omega_i^\vee$ *to* $w_o s_i \omega_i^\vee$ *in* $V(\omega_i^\vee)$,

where $V(\omega_i^\vee)$ is the i-th fundamental representation for the Langlands dual Lie algebra $\mathfrak{g}^\vee$.

References

[B_1] P. Belkale, Local systems on $\mathbb{P}^1 - S$ for S a finite set, *Compositio Math.* **129** (2001), 67–86.

[B_2] P. Belkale, Geometric proofs of Horn and saturation conjectures, *J. Alg. Geom.* **15** (2006), 133–173.

[B_3] P. Belkale, Geometric proof of a conjecture of Fulton, *Advances Math.* **216** (2007), 346–357.

[BK_1] P. Belkale and S. Kumar, Eigenvalue problem and a new product in cohomology of flag varieties, *Inventiones Math.* **166** (2006), 185–228.

[BK_2] P. Belkale and S. Kumar, Eigencone, saturation and Horn problems for symplectic and odd orthogonal groups, *J. of Algebraic Geom.* **19** (2010), 199–242.

[BCH] G. Benkart, M. Chakrabarti, T. Halverson, R. Leduc, C.Y. Lee and J. Stroomer, Tensor product representations of general linear groups and their connections with Brauer algebras, *J. of Algebra* **166** (1994), 529–567.

[BS] A. Berenstein and R. Sjamaar, Coadjoint orbits, moment polytopes, and the Hilbert-Mumford criterion, *Journ. Amer. Math. Soc.* **13** (2000), 433–466.

[BZ] A. Berenstein and A. Zelevinsky, Tensor product multilplicities, canonical bases and totally positive varieties, *Inventiones Math.* **143** (2001), 77–128.

[BT] A. Borel and J. Tits, Homomorphismes "abstraits" de groupes algébriques simples, *Ann. of Math.* **97** (1973), 499–571.

[Bot] R. Bott, Homogeneous vector bundles, *Ann. of Math.* **66** (1957), 203–248.

[Bo] N. Bourbaki, *Groupes et Algèbres de Lie*, Chap. 4–6, Masson, Paris, 1981.

[BrK] M. Brion and S. Kumar, *Frobenius Splitting Methods in Geometry and Representation Theory*, Birkhäuser, 2005.

[BD] T. Bröcker and T. tom Dieck, *Representations of Compact Lie Groups*, Springer-Verlag, 1985.

[BL] J. Brown and V. Lakshmibai, Wahl's conjecture for a minuscule G/P, *Proc. Indian Acad. Sci (Math. Sci.)* **119** (2009), 571–592.

[C] C. Chevalley, Classification des groupes de Lie algébriques, *Séminaire C. Chevalley 1956–58, vol. 1*, Ecole Normale Supérieure, 1958.

[DW] H. Derksen and J. Weyman, Semi-invariants of quivers and saturation for Littlewood-Richardson coefficients, *J. Amer. Math. Soc.* **13** (2000), 467–479.

[DR_1] I. Dimitrov and M. Roth, Cup products of line bundles on homogeneous varieties and generalized PRV components of multiplicity one, Preprint (2009).

[DR_2] I. Dimitrov and M. Roth, Geometric realization of PRV components and the Littlewood-Richardson cone, Preprint (2009).

[D] S. Donkin, *Rational Representations of Algebraic Groups*, Lecture Notes in Math. **1140**, Springer-Verlag, 1985.

[E] A.G. Èlashvili, Invariant algebras, In: *Lie Groups, their Discrete Subgroups, and Invariant Theory* (ed. E. B. Vinberg), Advances in Soviet Math. **8**, Amer. Math. Soc., Providence, 1992, 57–64.

[F_1] W. Fulton, *Intersection Theory, 2nd edn.*, Springer-Verlag, 1998.

[F_2] W. Fulton, Eigenvalues, invariant factors, highest weights, and Schubert calculus, *Bull. Amer. Math. Soc. (N.S.)* **37** (2000), 209–249.

[GP] S. Grimm and J. Patera, Decomposition of tensor products of the fundamental representations of E_8, *CRM Proc. Lecture Notes* **11** (1997), Amer. Math. Soc., 329–355.

[H] R. Hartshorne, *Algebraic Geometry*, Springer-Verlag, 1977.

[J] J.C. Jantzen, *Representations of Algebraic Groups, 2nd edn.*, Am. Math. Soc., 2003.

[Jo] A. Joseph, On the Demazure character formula, *Ann. Sci. Éc. Norm. Supér.* **18** (1985), 389–419.

[KKM] M. Kapovich, S. Kumar and J. J. Millson, The eigencone and saturation for Spin(8), *Pure and Applied Math. Quarterly* **5** (2009), 755–780.

[KLM] M. Kapovich, B. Leeb and J. J. Millson, Convex functions on symmetric spaces, side lengths of polygons and the stability inequalities for weighted configurations at infinity, *Journal of Differential Geometry* **81** (2009), 297–354.

[KM_1] M. Kapovich and J. J. Millson, Structure of the tensor product semigroup, *Asian J. of Math.* **10** (2006), 493–540.

[KM_2] M. Kapovich and J. J. Millson, A path model for geodesics in Euclidean buildings and its applications to representation theory, *Groups, Geometry and Dynamics* **2** (2008), 405-480.

[Ka] M. Kashiwara, On crystal bases, *Canadian Math. Soc. Conf. Proc.* **16** (1995), 155–197.

[Kas] S. Kass, Explicit decompositions of some tensor products of modules for simple complex Lie algebras, *Comm. in Algebra* **15** (1987), 2251–2261.

[Ki] F. Kirwan, *Cohomology of Quotients in Symplectic and Algebraic Geometry*, Princeton University Press, 1984.

[Kly] A. Klyachko, Stable bundles, representation theory and Hermitian operators, *Selecta Mathematica* **4** (1998), 419–445.

[KT] A. Knutson and 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), 1055–1090.

[KTW] A. Knutson, T. Tao and C. Woodward, The honeycomb model of $\mathrm{GL}_n(\mathbb{C})$ tensor products II: Puzzles determine facets of the Littlewood-Richardson cone, *J. Amer. Math. Soc.* **17** (2004), 19–48.

[Koi] K. Koike, On the decomposition of tensor products of the representations of the classical groups: By means of the universal characters, *Advances in Math.* **74** (1989), 57–86.

[Ko] B. Kostant, A formula for the multiplicity of a weight, *Trans. Am. Math. Soc.* **93** (1959), 53–73.

[K_1] S. Kumar, Proof of the Parthasarathy-Ranga Rao-Varadarajan conjecture, *Inventiones Math.* **93** (1988), 117–130.

[K_2] S. Kumar, A refinement of the PRV conjecture, *Inventiones Math.* **97** (1989), 305–311.

[K_3] S. Kumar, Proof of Wahl's conjecture on surjectivity of the Gaussian map for flag varieties, *Amer. J. Math.* **114** (1992), 1201–1220.

[K_4] S. Kumar, *Kac-Moody Groups, their Flag Varieties and Representation Theory*, Progress in Mathematics, vol. **204**, Birkhäuser, 2002.

[KuLM] S. Kumar, B. Leeb and J. J. Millson, The generalized triangle inequalities for rank 3 symmetric spaces of noncompact type, *Contemp. Math.* **332** (2003), 171–195.

[KS] S. Kumar and J. Stembridge, Special isogenies and tensor product multiplicities, *Inter. Math. Res. Not.*, vol. **2007** (no. 20) (2007), 1–13.

[LRS] V. Lakshmibai, K.N. Raghavan and P. Sankaran. Wahl's conjecture holds in odd characteristics for symplectic and orthogonal Grassmannians, *Central European J. of Math.* **7** (2009), 214–223.

[LCL] M.A. A. van Leeuwen, A.M. Cohen and B. Lisser, LiE, *A package for Lie group computations*, Computer Algebra Nederland, 1992.

[L_1] P. Littelmann, A generalization of the Littlewood-Richardson rule, *J. of Algebra* **130** (1990), 328–368.

[L_2] P. Littelmann, A Littlewood-Richardson rule for symmetrizable Kac-Moody algebras, *Invent. Math.* **116** (1994), 329–346.

[L_3] P. Littelmann, Paths and root operators in representation theory, *Ann. of Math.* **142** (1995), 499–525.

[Lu] G. Lusztig, Canonical bases arising from quantized enveloping algebras. II, *Prog. Theor. Phys.* **102** (1990), 175–201.

[Ma] I.G. Macdonald, *Symmetric Functions and Hall Polynomials, 2nd edn.*, Oxford Mathematical Monographs, 1995.

[M_1] O. Mathieu, Construction d'un groupe de Kac-Moody et applications, *Compositio Math.* **69** (1989), 37–60.

[M_2] O. Mathieu, Filtrations of G-modules, *Ann. Sci. Éc. Norm. Supér.* **23** (1990), 625–644.

[MP] V.B. Mehta and A.J. Parameswaran, On Wahl's conjecture for the Grassmannians in positive characteristic, *Internat. J. Math.* **8** (1997), 495–498.

[MMP] W.G. McKay, R.V. Moody and J. Patera, Decomposition of tensor products of E_8 representations, *Algebras, Groups, and Geometries* **3** (1986), 286–328.

[MFK] D. Mumford, J. Fogarty and F. Kirwan, *Geometric Invariant Theory, 3rd edn.*, Ergebnisse der Mathematik und ihrer grenzgebiete, vol. **34**, Springer, 1994.

[PRV] K.R. Parthasarathy, R. Ranga Rao, and V.S. Varadarajan, Representations of complex semi-simple Lie groups and Lie algebras, *Ann. of Math.* **85** (1967), 383–429.

[PS] K. Purbhoo and F. Sottile, The recursive nature of cominuscule Schubert calculus, *Advances Math.* **217** (2008), 1962–2004.

[Ra] K.N. Rajeswari, Standard monomial theoretic proof of PRV conjecture, *Communications in Algebra* **19** (1991), 347–425.

[RR] S. Ramanan and A. Ramanathan, Some remarks on the instability flag, *Tohoku Math. J.* **36** (1984), 269–291.

[R_1] N. Ressayre, Geometric invariant theory and the generalized eigenvalue problem, *Inventiones Math.* (2010).

[R_2] N. Ressayre, A short geometric proof of a conjecture of Fulton, Preprint (2009).

[R_3] N. Ressayre, Eigencones and the PRV conjecture, Preprint (2009).

[ReR] N. Ressayre and E. Richmond, Branching Schubert calculus and the Belkale-Kumar product on cohomology, Preprint (2009).

[Ri_1] E. Richmond, A partial Horn recursion in the cohomology of flag varieties, J. Alg. Comb. **30** (2009), 1–17.

[Ri_2] E. Richmond, A multiplicative formula for structure constants in the cohomology of flag varieties, Preprint (2008).

[Sj] R. Sjamaar, Convexity properties of the moment mapping re-examined, *Advances Math.* **138** (1998), 46–91.

[S] R. Steinberg. A general Clebsch-Gordan theorem, *Bull. Amer. Math. Soc.* **67** (1961), 406–407.

[St_1] J. R. Stembridge, Computational aspects of root systems, Coxeter groups, and Weyl characters, In: *Interaction of Combinatorics and Representation Theory*, MSJ Memoirs **11**, Math. Soc. Japan (2001), 1–38.

[St_2] J. R. Stembridge, Combinatorial models for Weyl characters, *Advances Math.* **168** (2002), 96–131.

[W] J. Wahl, Gaussian maps and tensor products of irreducible representations, *Manuscripta Math.* **73** (1991), 229–259.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Some Applications of the Trace Formula and the Relative Trace Formula

Erez M. Lapid*

Abstract

The trace formula is a major tool in the theory of automorphic forms. It was conceived by Selberg and extensively developed by Arthur. Among other things it is applicable to the study of spectral asymptotics as well as to (special cases of) Langlands functoriality conjectures. An important variant invented by Jacquet – the relative trace formula – is used to study period integrals and invariant functionals.

Mathematics Subject Classification (2010). Primary 11F72; Secondary 11F70, 58C40.

Keywords. Trace formula

Introduction

The trace formula was introduced by Selberg in his seminal paper [51], and was mostly studied by him in the context of finite volume quotients of the hyperbolic plane. Among its early applications are the Weyl law for quotients by congruence subgroups and the prime geodesic theorem. Later on, it was conceived by Langlands that the trace formula can be extremely useful to his functoriality conjectures (spurred on by works of Shimizu, Saito and Shintani [53, 44, 55]). Up until today, this theme continues to be at the heart of most attacks (and results) on functoriality. Arthur's groundbreaking development of the trace formula from its initial non-invariant form to its final stable form is much driven by the functoriality conjectures. In particular, a great emphasis is put on being able to compare trace formulas on different groups, and to manipulate the terms to be invariant (and then stably invariant). On the other

*Author partially supported by grants from the German-Israeli-Foundation, United States-Israel Binational Science Foundation and the Israel Science Foundation.

Einstein Institute of Mathematics, The Hebrew University of Jerusalem,Givat Ram, Jerusalem 91904, Israel. E-mail: erezla@math.huji.ac.il.

hand, in order to extend earlier results of the trace formula to the asymptotic spectral analysis of locally symmetric spaces, it suffices to work with the trace formula in its non-invariant form, but it is necessary to study analytically the various distributions of the trace formula.

In the first part we will describe some recent progress on the structure of the trace formula and its applications to spectral asymptotics. For the applications of the trace formula to functoriality and the pioneering work of Ngô on the *Fundamental Lemma* we refer to the second part of the beautiful survey [8] and to Ngô's article in these proceedings.

In the second part we turn to the relative trace formula, which is a variant of the trace formula invented by Jacquet. The main motivation comes from the study of period integrals. We will focus on the case of unitary groups inside the general linear group of a quadratic extension.

It is a pleasure to thank Hervé Jacquet, Jonathan Rogawski, Peter Sarnak, Akshay Venkatesh and my coauthors Tobias Finis, Werner Müller and Omer Offen for fruitful discussions over the years.

1. Arthur's Trace Formula

We will give a brief description here, freely quoting Arthur's results. We refer to [8] and the references cited therein for more details. However, our presentation will be somewhat different.

1.1. Notation. Let G be a connected reductive group over $\mathbb{Q}$. Fix a maximal split torus T_0 of G. Any parabolic subgroup of G containing T_0 and defined over $\mathbb{Q}$ admits a unique Levi part (defined over $\mathbb{Q}$) containing T_0. Denote by $\mathcal{L}$ the set of Levi subgroups obtained this way. In particular the centralizer M_0 of T_0 in G is the minimal element of $\mathcal{L}$. For any $M \in \mathcal{L}$ let $\mathcal{P}(M)$ (resp. $\mathfrak{F}(M)$) be the sets of parabolic subgroups of G defined over $\mathbb{Q}$ with Levi part M (resp. containing M).

For any connected algebraic group L over $\mathbb{Q}$ let $X^*(L)$ be the lattice of characters of L defined over $\mathbb{Q}$ (factoring through the Levi part of L) and set $\mathfrak{a}_L^* = X^*(L) \otimes \mathbb{R}$. Let $\mathfrak{a}_L$ be the dual space and define $H = H_L : L(\mathbb{A}) \to \mathfrak{a}_L$ by $\langle \chi, H(l) \rangle = \log|\chi(l)|$ for any $\chi \in X^*(L)$ (extended to a character $L(\mathbb{A}) \to \mathbb{A}^*$). Also, let T_L be the split part of the center of the Levi part of L and let $A_L = T_L(\mathbb{R})^0$ so that $A_L \simeq \mathfrak{a}_L$ via H_L. For $M, L \in \mathcal{L}$ with $M \subseteq L$ the restriction map embeds $\mathfrak{a}_L^*$ in $\mathfrak{a}_M^*$ and we set $(\mathfrak{a}_M^L)^* = \mathfrak{a}_M^*/\mathfrak{a}_L^*$. Fix a suitable maximal compact $\mathbf{K}$ of $G(\mathbb{A})$ and for any $P \in \mathfrak{F}(M_0)$ extend the function $H_P : P(\mathbb{A}) \to \mathfrak{a}_P$ to a right $\mathbf{K}$-invariant function on $G(\mathbb{A})$.

1.2. Regularization. For a general $f \in C_c^\infty(G(\mathbb{A}))$ the kernel

$$K_f(x,y) = \sum_{\gamma \in G(\mathbb{Q})} \int_{A_G} f(x^{-1} z \gamma y)\, dz$$

of the integral operator

$$R(f)\varphi(x)=\int_{G(\mathbb{A})}f(g)\varphi(xg)\,dg$$

on $L^2(A_GG(\mathbb{Q})\backslash G(\mathbb{A}))$ is not integrable over the diagonal. In order to regularize the trace formula fix $P_0\in\mathcal{P}(M_0)$. Let $\mathcal{C}_+$ be the cone in $\mathfrak{a}_0:=\mathfrak{a}_{P_0}$ spanned by $\mathfrak{a}_G$ and the coroots of T_0 on the unipotent radical of P_0. For T in the positive Weyl chamber $\mathfrak{a}_{0,+}$ of $\mathfrak{a}_0$ let

$$\mathfrak{S}_{\leq T}=\{g\in G(\mathbb{A}):T-H_{P_0}(\gamma g)\in\mathcal{C}_+\text{ for all }\gamma\in G(\mathbb{Q})\}.^1$$

By reduction theory $A_GG(\mathbb{Q})\backslash\mathfrak{S}_{\leq T}$ is compact and

$$\int_{A_GG(\mathbb{Q})\backslash\mathfrak{S}_{\leq T}}K_f(x,x)\,dx$$

approximates a polynomial $P^T(f)$ in T of degree $\dim\mathfrak{a}_0^G$ with an exponentially decreasing error term as $T\to\infty$ in any closed subcone of $\mathfrak{a}_{0,+}$. Introduce the equivalence relation $\sim$ on $G(\mathbb{Q})$ by $\gamma_1\sim\gamma_2$ if they are conjugate in $G(\overline{\mathbb{Q}})$ and their semisimple parts are conjugate in $G(\mathbb{Q})$. The polynomial approximation above is compatible with the decomposition of K_f according to these equivalence classes in the sense that for any $\sim$-class $[\gamma]$ there exists a polynomial $P^T_{[\gamma]}(f)$ such that

$$\sum_{[\gamma]}\left|\int_{A_GG(\mathbb{Q})\backslash\mathfrak{S}_{\leq T}}\sum_{\delta\in[\gamma]}f(x^{-1}\delta x)\,dx-P^T_{[\gamma]}(f)(T)\right|$$

is exponentially small in T for $T\to\infty$ as above. In particular, $P^T(f)=\sum P^T_{[\gamma]}(f)$. The regularized trace $J(f)$ of $R(f)$ is by definition the value of this polynomial at a suitable point T_0 (depending on the choice of $\mathbf{K}$). Setting $J_{[\gamma]}(f)=P^{T_0}_{[\gamma]}(f)$ one gets

$$J(f)=\sum_{[\gamma]}J_{[\gamma]}(f).$$

1.3. Geometric side. Suppose that $[\gamma]$ is semisimple. Let C_γ be the centralizer of γ in G and let C_γ^0 be its identity component. Let M be the centralizer of the split part of $Z(C_\gamma^0)$, i.e. the smallest Levi subgroup of G containing C_γ^0. Then γ is elliptic in M and by conjugating γ we may assume

[1] The exact shape of $\mathfrak{S}_{\leq T}$ is probably not too important. Other families of exhausting domains for $A_GG(\mathbb{Q})\backslash G(\mathbb{A})$, under mild conditions, would do too.

that M is standard. Then

$$J_{[\gamma]}(f) = \int_{A_M C_\gamma(\mathbb{Q})\backslash G(\mathbb{A})} f(g^{-1}\gamma g)\,\mathrm{vol}(\mathfrak{P}_M(g))\;dg$$
$$= \frac{\mathrm{vol}(A_M C^0_\gamma(\mathbb{Q})\backslash C^0_\gamma(\mathbb{A}))}{[C_\gamma(\mathbb{Q}) : C^0_\gamma(\mathbb{Q})]} \int_{C^0_\gamma(\mathbb{A})\backslash G(\mathbb{A})} f(g^{-1}\gamma g)\,\mathrm{vol}(\mathfrak{P}_M(g))\;dg$$

where $\mathfrak{P}_M(g)$ is the convex hull in $\mathfrak{a}^G_M$ of $\{H_P(g) : P \in \mathcal{P}(M)\}$. We note that when the $H_P(g)$'s are distinct, the faces of $\mathfrak{P}_M(g)$ correspond to $\mathfrak{F}(M)$ – the face corresponding to Q is the convex hull of $\{H_P(g) : P \subseteq Q\}$.

In particular, the elliptic contribution is

$$\sum_{[\gamma]\in G(\mathbb{Q})\ \text{elliptic}} \int_{A_G C_\gamma(\mathbb{Q})\backslash G(\mathbb{A})} f(g^{-1}\gamma g)\;dg$$
$$= \sum_{[\gamma]\in G(\mathbb{Q})\ \text{elliptic}} \frac{\mathrm{vol}(A_G C^0_\gamma(\mathbb{Q})\backslash C^0_\gamma(\mathbb{A}))}{[C_\gamma(\mathbb{Q}) : C^0_\gamma(\mathbb{Q})]} \int_{C^0_\gamma(\mathbb{A})\backslash G(\mathbb{A})} f(g^{-1}\gamma g)\;dg.$$

The contribution from the non-semisimple conjugacy classes is more complicated. Arthur reduces it to the unipotent case and analyzes the ensuing local distributions [7, 5, 6]. However, his method does not seem to give an effective way to compute the global coefficients, and some control over them is essential for certain applications of the trace formula. In the case of GL_2 the non-trivial coefficient is Euler's γ constant. More generally, in the rank one case, one can write the coefficients above in terms of the coefficients of the Laurent expansion of zeta functions of prehomogeneous spaces, studied by Sato, Shintani, and others [26, 48].

Let us describe work in progress with Tobias Finis aiming to explicate this relation in higher rank. Let $\mathfrak{o}$ be a geometric unipotent orbit in G defined over $\mathbb{Q}$ and let $P = MU$ be the standard Jacobson-Morozov parabolic associated to $\mathfrak{o}$. Its unipotent radical U is equipped with a descending filtration $U^{\geq i}$ associated to the action of the Jacobson-Morozov torus and $U^{\geq 2,\mathrm{reg}} = U^{\geq 2} \cap \mathfrak{o}$ is a principal open set of $U^{\geq 2}$. Using Iwasawa decomposition one gets (assuming f is $\mathrm{Ad}\,\mathbf{K}$-invariant, which we may)

$$j^T_{\mathfrak{o}}(f) := \int_{A_G G(Q)\backslash \mathfrak{S}_{\leq T}} \sum_{\gamma\in\mathfrak{o}(\mathbb{Q})} f(x^{-1}\gamma x)\;dx$$
$$= \int_{A_G M(\mathbb{Q})\backslash M(\mathbb{A})} \sum_{u\in U^{2,\mathrm{reg}}(\mathbb{Q})} \int_{C_U(u)(\mathbb{A})\backslash U(\mathbb{A})} \left[\int_{C_U(u)(\mathbb{Q})\backslash C_U(u)(\mathbb{A})} \mathbf{1}_{\mathfrak{S}_{\leq T}}(v'vm)dv'\right]$$
$$f(m^{-1}v^{-1}uvm)\delta_U(m)^{-1}\;dv\;dm$$

where $U^{2,\mathrm{reg}} = U^{\geq 2,\mathrm{reg}}/U^{>2}$ is the regular part of the prehomogeneous vector space $U^2 = U^{\geq 2}/U^{>2}$ with respect to M

The key point is that we can approximate the integral in brackets independently of v by the characteristic function $\tilde{F}^M(m,T)$ of the set

$$\{m \in M(\mathbb{A}) : T - H_{P_0}(\gamma m) \in \mathcal{C}_+ \text{ for all } \gamma \in M(\mathbb{Q})\}.$$

Therefore, we can approximate $j^T_{\mathfrak{o}}(f)$ by

$$\int_{A_G M(\mathbb{Q})\backslash M(\mathbb{A})} \tilde{F}^M(m,T) \sum_{u \in U^{2,\mathrm{reg}}(\mathbb{Q})} f_{U^{>2}}(m^{-1}um)\delta_{U^{\leq 2}}(m)^{-1}\, dm,$$

where $\delta_{U^{\leq 2}}$ is the modulus function of $MU^{\leq 2}$ and we put

$$f_{U^{>2}}(u) = \int_{U^{>2}(\mathbb{A})} f(uv)\, dv.$$

In the simplest cases we can replace the above by the term $j^T_{\mathfrak{o},\,\mathrm{main}}(f)$ given by

$$\int_{A_G M(\mathbb{Q})\backslash M(\mathbb{A})} \hat{\tau}_P(T - H_M(m)) \sum_{u \in U^{2,\mathrm{reg}}(\mathbb{Q})} f_{U^{>2}}(m^{-1}um)\delta_{U^{\leq 2}}(m)^{-1}\, dm,$$

where $\hat{\tau}_P$ is the characteristic function of the cone spanned by $\mathfrak{a}_0^M$ and the positive coroots. The subspace $\mathfrak{a}_C = H_M(C_u(\mathbb{A})) \subseteq \mathfrak{a}_M$ is independent of u and the orthogonal complement $\mathfrak{a}_C^\perp \subseteq \mathfrak{a}_M^*$ is spanned by the set X of fundamental characters for the M-action on U^2.

In the cases at hand the zeta function

$$Z(\phi,\lambda) = \int_{\mathfrak{a}_M/\mathfrak{a}_C} e^{-\langle \lambda, X\rangle} \int_{M(\mathbb{Q})\backslash M(\mathbb{A})^1} \sum_{u \in U^{2,\mathrm{reg}}(\mathbb{Q})} \phi(e^{-X}m^{-1}ume^X)\, dm\, dX$$

converges for $\lambda \in \mathfrak{a}_{C,\mathbb{C}}^\perp$ such that the coordinates μ_χ of $\lambda - \rho_{M,2}$ with respect to X are on the right half-plane [45], and admits meromorphic continuation for $\operatorname{Re}\mu_\chi > -\varepsilon$ for some $\varepsilon > 0$ with only simple poles along the hyperplanes $\mu_\chi = 0$.

Let

$$\theta_P(\lambda) = \operatorname{vol}\left(\mathfrak{a}_P^G / \sum_{\alpha^\vee \in \Delta_P^\vee} \mathbb{Z}\alpha^\vee\right) \prod_{\alpha^\vee \in \Delta_P^\vee} \langle \lambda, \alpha^\vee\rangle$$

where $\Delta_P^\vee$ is the set of simple coroots of P. Applying Mellin inversion, we obtain for $\lambda_0 \in \mathfrak{a}_C^\perp \cap (\mathfrak{a}_P^*)_+$

$$j^T_{\mathfrak{o},\,\mathrm{main}}(f) = \int_{\lambda \in \mathfrak{a}_{C,\mathbb{C}}^\perp :\, \operatorname{Re}\lambda = \lambda_0} \frac{Z(f_{U^{>2}}, \lambda + \rho_{M,\leq 2})}{\theta_P(\lambda)} e^{\langle \lambda, T\rangle}\, d\lambda.$$

Assume for simplicity that $\mathfrak{o}$ is *even*. Let $\hat{\theta}_X(\lambda)$ be the product of the coordinates of $\lambda \in \mathfrak{a}_{C,\mathbb{C}}^\perp$ with respect to X. Writing $z(\lambda) = \hat{\theta}_X(\lambda) Z(f_{U^{>2}}, \lambda + \rho_{M,\leq 2})$ and

$$z(\lambda) = \int_{\mathfrak{a}_M/\mathfrak{a}_C} \psi(x) e^{\langle \lambda, x\rangle}\, dx, \quad \lambda \in \mathfrak{a}_C^\perp,$$

we have

$$\int_{\lambda\in\mathfrak{a}_{C,\mathbb{C}}^{\perp},\,\mathrm{Re}\lambda=\lambda_0}\frac{z(\lambda)e^{\langle\lambda,T\rangle}}{\hat{\theta}_X(\lambda)\theta_P(\lambda)}\,d\lambda=\int_{\mathfrak{a}_M/\mathfrak{a}_C}\psi(x)v(x+T)\,dx$$

for the volume function

$$v(x)=\mathrm{vol}(\{H\in\mathfrak{a}_M\mid\hat{\tau}_P(x-H)=1\text{ and }\langle\chi,H\rangle\geq 0,\chi\in X\}).$$

For instance, in the case where $\mathfrak{o}$ is the unipotent orbit of type $(m,\dots,m)$ in GL_n (in which $X=\Delta_P$ and $Z(\lambda)$ is essentially a product of Riemann zeta functions) we get

$$P_{\mathfrak{o}}^T(f)=\lim_{\lambda\to 0}\frac{1}{|W(M)|}\sum_{w\in W(M)}\frac{z(w\lambda)e^{\langle w\lambda,T\rangle}}{\theta_P(w\lambda)}$$

where $W(M)=N_G(M)/M$. In other cases there are additional terms beside $j_{\mathfrak{o},\,\mathrm{main}}^T$ coming from proper parabolic subgroups of P containing the identity component of the centralizer of some $u\in U^{\geq 2,\mathrm{reg}}(\mathbb{Q})$. Moreover, the zeta function has to be regularized in the cases of *incomplete type* where $C_u(\mathbb{Q})\backslash C_u(\mathbb{A})\cap P(\mathbb{A})^1$ has infinite volume for some $u\in U^{\geq 2,\mathrm{reg}}(\mathbb{Q})$.

Before we turn to the spectral side we consider the following situation. Suppose that $\mathfrak{P}\subseteq\mathbb{R}^d$ is a d-dimensional polytope, i.e. the convex hull of finitely many points in $\mathbb{R}^d$ such that $\mathfrak{P}-\mathfrak{P}$ spans $\mathbb{R}^d$. For each face F of $\mathfrak{P}$ (not necessarily maximal) fix a point $v(F)\in F$. For any "flag" $\mathfrak{f}: F_0\subseteq\ldots\subseteq F_d=\mathfrak{P}$ where F_i is an i-dimensional face of $\mathfrak{P}$ let $\Delta(\mathfrak{f})$ be the convex hull of $\{v(F_0),\dots,v(F_d)\}$. Let $\mathfrak{F}$ be the set of flags. Then $\Delta(\mathfrak{f}),\mathfrak{f}\in\mathfrak{F}$ is a decomposition of $\mathfrak{P}$ into simplices with pairwise disjoint interiors. (For instance, if $v(\mathfrak{P})$ is the barycenter of $\mathfrak{P}$ then we get the barycentric subdivision.) In particular,

$$\mathrm{vol}\,\mathfrak{P}=\sum_{\mathfrak{f}\in\mathfrak{F}}\mathrm{vol}\,\Delta(\mathfrak{f}).\tag{1}$$

1.4. The spectral side. The point of departure for the spectral decomposition of Arthur's trace formula is Langlands' description of the decomposition of $L^2(G(\mathbb{Q})\backslash G(\mathbb{A}))$ in terms of the discrete spectrum of Levi subgroups [34]. For any $P\in\mathcal{P}(M)$ let

$$\mathcal{A}_P^2=\mathrm{Ind}_{P(\mathbb{A})}^{G(\mathbb{A})}L^2_{\mathrm{disc}}(A_MM(\mathbb{Q})\backslash M(\mathbb{A})).$$

On this space there is a family of induced representations $I_P(\lambda)$, $\lambda\in\mathfrak{a}_P^*$. The theory of Eisenstein series gives rise to intertwining maps from $I_P(\lambda)$ to the space of automorphic forms on $G(F)\backslash G(\mathbb{A})$ which furnish the spectral decomposition of $L^2(G(F)\backslash G(\mathbb{A}))$. Alongside it provides a family of unitary intertwining operators

$$\mathcal{F}_{Q|P}(\lambda):\mathcal{A}_P^2\to\mathcal{A}_Q^2\quad P,Q\in\mathcal{P}(M),\lambda\in\mathfrak{ia}_M^*$$

satisfying $I_Q(\lambda)\circ\mathcal{F}_{Q|P}(\lambda)=\mathcal{F}_{Q|P}(\lambda)\circ I_P(\lambda)$ and the functional equations

$$\mathcal{F}_{R|Q}(\lambda)\circ\mathcal{F}_{Q|P}(\lambda)=\mathcal{F}_{R|P}(\lambda)\quad \lambda\in\mathrm{i}\mathfrak{a}_M^* \text{ for any } P,Q,R\in\mathcal{P}(M).$$

Moreover, as a function of λ, $\mathcal{F}_{Q|P}(\lambda)$ is invariant under translation by the span of $\mathrm{i}\overline{\mathfrak{a}_{P,+}^*}\cap\mathrm{i}\overline{\mathfrak{a}_{Q,+}^*}$ (where $\overline{\mathfrak{a}_{P,+}^*}$ is the closure of the positive Weyl chamber of $\mathfrak{a}_P^*$).

Note that in the non-compact case already the fact that $R_{\mathrm{disc}}(f)$ is of trace class is by no means obvious, even using Langlands' description of the discrete spectrum in terms of residues of Eisenstein series. It was proved by Müller [42].

Fix $P\in\mathcal{P}(M)$ and $L\in\mathcal{L}$ containing M. Let $k=\dim\mathfrak{a}_M^L$. For any $Q\in\mathfrak{F}(L)$ choose $S(Q)\in\mathcal{P}(M)$ which is contained both in Q and in an element of $\mathcal{P}(L)$ (i.e., L is standard with respect to $S(Q)$). For any chain of parabolic subgroups $\mathfrak{f}: Q_0\subseteq\ldots\subseteq Q_k=G$ such that $Q_0\in\mathcal{P}(L)$ and Q_{i-1} is maximal in Q_i, $i=1,\ldots,k$ set

$$\partial_{\mathfrak{f}}\mathcal{F}(\lambda)=\mathrm{vol}((\mathfrak{a}_L^G)^*/\mathbb{Z}\lambda_0+\cdots+\mathbb{Z}\lambda_{k-1})^{-1}\mathcal{F}_{P|S(Q_0)}(\lambda)$$
$$\partial_{\lambda_0}\mathcal{F}_{S(Q_0)|S(Q_1)}(\lambda)\,\partial_{\lambda_1}\mathcal{F}_{S(Q_1)|S(Q_2)}(\lambda)\cdots\partial_{\lambda_{k-1}}\mathcal{F}_{S(Q_{k-1})|S(G)}(\lambda)\,\mathcal{F}_{S(G)|P}(\lambda)$$

where $\lambda_i\in\mathfrak{a}_{Q_i,+}^*$ and ∂_{λ_i} denotes the (first-order) directional derivative. This operator is defined on a dense subspace of $\mathcal{A}_P^2$ and does not depend on the choice of the λ_i's. A follow up of Müller's work on the trace class conjecture [41] yields the convergence of

$$\int_{\mathrm{i}(\mathfrak{a}_L^G)^*}\|\partial_{\mathfrak{f}}\mathcal{F}(\lambda)I_P(f,\lambda)\|_{1,\mathcal{A}_P^2}\,d\lambda \tag{2}$$

where $\|\cdot\|_1$ denotes the trace norm. In fact, fix any compact open subgroup K of $G(\mathbb{A}_{\mathrm{fin}})$ and consider the space $\mathcal{F}(G(\mathbb{A});K)$ of right-K-invariant functions on $G(\mathbb{A})$ such that

$$\|f\star X\|_{L^1(G(\mathbb{A}))}<\infty$$

for all $X\in\mathcal{U}(\mathfrak{g})$. Then (2) is a continuous seminorm on $\mathcal{F}(G(\mathbb{A});K)$. An analogous statement for the geometric side holds at least for the elliptic contribution [21].

The sum

$$\mathcal{D}_{P,L}\mathcal{F}(\lambda)=\frac{1}{k!}\sum\partial_{\mathfrak{f}}\mathcal{F}(\lambda)$$

over all $\mathfrak{f}$ as above is independent of the choice of $S(Q)$'s. For $L=M$ this expression is analogous to the right-hand side of (1) for the polytope $\mathfrak{P}_M(g)$ where for the face F corresponding to Q we take $v(F)$ to be the vertex $H_{S(Q)}(g)$ of F.

For $M\in\mathcal{L}$ let $W(M)=N_M(\mathbb{Q})/M(\mathbb{Q})$ where N_M is the normalizer of M in G. We can identify $W(M)$ with a subgroup of the Weyl group $W(M_0)$. For any $s\in W(M)$ let M_s be the smallest subgroup in $\mathcal{L}$ containing M and s. Thus $\mathfrak{a}_{M_s}$ is the fixed points of s on $\mathfrak{a}_M$. We denote by $\mathfrak{c}_s$ the conjugation operator $\mathfrak{c}_s:\mathcal{A}_P^2\to\mathcal{A}_{sPs^{-1}}^2$.

Theorem 1.1 ([22]). *The spectral side of Arthur's trace formula is given by*

$$\sum_{[P]} \frac{1}{|W(M)|} \sum_{s \in W(M)} |\det(s-1|\mathfrak{a}_M^{M_s})|^{-1} \int_{\mathrm{i}(\mathfrak{a}_{M_s}^G)^*} \mathrm{tr}(\mathcal{D}_{P,M_s} \mathcal{F}(\lambda) \mathcal{F}_{P|sPs^{-1}}(\lambda) \mathfrak{c}_s I_P(f,\lambda))_{\mathcal{A}_P^2} \, d\lambda \quad (3)$$

where $P \in \mathfrak{F}(M_0)$ runs over a set of representatives of associated classes of parabolic subgroups and $M \in \mathcal{L}$ is the Levi part of P.

The expression above is an explicit version of Arthur's spectral expansion [4]. It is absolutely convergent (with respect to the trace norm) even for $f \in \mathcal{F}(G(\mathbb{A}); K)$.

1.5. The Weyl law. One of Selberg's early applications of the trace formula for hyperbolic surfaces is showing that Maass forms exist in abundance. Namely, let Δ be the Laplacian on the hyperbolic plane $\mathbb{H}$ and let Γ be a congruence subgroup of $\mathrm{SL}_2(\mathbb{R})$. Denote by N_T the counting function for the number of linearly independent solutions of $(\Delta + \lambda)f = 0$ in $L^2(\Gamma\backslash\mathbb{H})$ with $\lambda < \frac{1}{4} + T^2$. Then

$$N_T = \frac{\mathrm{Area}(\Gamma\backslash\mathbb{H})}{4\pi} T^2 + O(T \log T).$$

The main point is that for a suitable family of test functions, the main contribution in the geometric side arises from the identity element, and on the spectral side from the discrete spectrum. The last point is subtle because it is based on the explicit computation of the determinant of the scattering matrix in terms of Dirichlet L-functions and the fact that the latter are entire functions of order one. It is generally not expected to hold in the non-arithmetic case [40] – see also [47].

In the case of compact surfaces, or even for a general compact Riemannian manifold M of dimension d the estimate

$$N_T = c_d \, \mathrm{vol}(M) T^d + O(T^{d-1}),$$

where c_d is a constant depending only on d, is classical. (Weyl considered a similar problem in the Euclidean plane.) In fact a general result for elliptic pseudo-differential operators is by now standard [27].

One may consider higher rank locally symmetric spaces $M = \Gamma\backslash G/K$ where Γ is a lattice in a semisimple group G and K is the maximal compact subgroup of G. The correct upper bound for the *cuspidal* spectrum was obtained in [18]. In the other direction, when Γ is a congruence subgroup, the correct lower bound for the cuspidal spectrum was obtained in [39] by a clever application

of a simple (and new) form of the trace formula.[2] It is interesting to point out however that as of now we do not have in general the correct upper bound for the *discrete* spectrum (even in the case of congruence subgroups).

In higher rank there is more than one invariant differential operator and it makes sense to ask about the joint distribution of eigenvalues. This can be rephrased representation-theoretically as follows. Let A be the identity component of a maximal split torus of G. The spherical representations of G are parameterized by the W-orbits of quasi-characters of A, where W is the Weyl group, and the *tempered* ones correspond to unitary characters. Let $m(\lambda)$ be the multiplicity of the spherical representation with parameter λ in $L^2(\Gamma\backslash G)$. Let μ_{pl} be the Plancherel measure on the vector space $\hat{A}$ of unitary characters of A. It is absolutely continuous with respect to the Lebesgue measure, and the density is given explicitly by the Gindikin-Karpelevic formula [25].

Theorem 1.2 (Duistermaat-Kolk-Varadarajan [19]). *Suppose that Γ is uniform (i.e., $\Gamma\backslash G$ is compact) and torsion free. Let $\Omega \subseteq \hat{A}$ be a bounded W-invariant domain with piecewise C^2-boundary. Then*

$$\sum_{\lambda\in W\backslash t\Omega} m(\lambda) = \frac{\mathrm{vol}(M)}{|W|}\mu_{\mathrm{pl}}(t\Omega) + O(t^{d-1}) \qquad \text{as } t \to \infty.$$

where $d = \dim G/K$. On the other hand,

$$\sum_{\lambda \text{ non-unitary}:\|\lambda\|<R} m(\lambda) = O(R^{d-2}).$$

In order to extend this kind of result to the non-compact case (possibly with an additional power of $\log t$ in the error term) one confronts the contribution of the continuous spectrum. The main issue is the control of the logarithmic derivatives of the co-rank one intertwining operators. For the general linear group the latter are fairly well understood. On the geometric side if Γ is sufficiently small the only contribution is from the unipotent conjugacy classes and one has to show that all but the trivial class are negligible. In fact, Arthur's description of the local distributions suffices and it is not necessary to know anything about the global coefficients. As a result one obtains the analogue of Theorem 1.2 (with a slightly weaker error term) for $G = \mathrm{SL}_n$ and Γ contained in a principal congruence subgroup of level 3 or higher [35]. A variant of this for other K-types is obtainable along the same lines using a suitable Paley-Weiner Theorem [12].

It is of import to extend this to estimate the traces of Hecke operators in certain families. This will have applications to the distribution of low lying zeros of L-functions. So far this has been carried out in rank one where the trace formula is known very explicitly [29]. To extend this to higher rank requires a better understanding of the terms in the trace formula (especially the global constants).

[2] Technically only a special case is considered, but this is not essential for the method.

1.6. Limit multiplicities. In this section I will describe a joint work in progress with Tobias Finis and Werner Müller. Let $\Gamma_1, \Gamma_2, \dots$ be a decreasing sequence of lattices of G such that Γ_n is normal in Γ_1 and $\cap\Gamma_n = 1$. Let μ_n be the atomic measure on $\hat{G}$ defined by the discrete spectrum of $L^2(\Gamma_n\backslash G)$, that is $\mu_n = \sum_{\pi\in\hat{G}} m(\pi)\delta_\pi$ where $m(\pi) = \dim \mathrm{Hom}_G(\pi, L^2(\Gamma_n\backslash G))$. Let μ_{pl} be the Plancherel measure on $\hat{G}$ that is

$$f(1) = \int_{\hat{G}} \mathrm{tr}\,\pi(f)\; d\mu_{\mathrm{pl}}(\pi).$$

By Harish-Chandra the support of μ_{pl} is the tempered spectrum $\hat{G}_{\mathrm{temp}}$ of G. One expects that in many circumstances the discrete spectrum of $L^2(\Gamma_n\backslash G)$ "tends to" the spectrum of $L^2(G)$, i.e. that $\frac{\mu_n}{\mathrm{vol}(\Gamma_n\backslash G)} \to \mu_{\mathrm{pl}}$. More precisely, for any $A \subseteq \hat{G}_{\mathrm{temp}}$ with $\mu_{\mathrm{pl}}(\partial A) = 0$ (the boundary in $\hat{G}_{\mathrm{temp}}$) we have $\frac{\mu_n(A)}{\mathrm{vol}(\Gamma_n\backslash G)} \to \mu_{\mathrm{pl}}(A)$ while $\frac{\mu_n(A)}{\mathrm{vol}(\Gamma_n\backslash G)} \to 0$ for any bounded $A \subseteq \hat{G} \setminus \hat{G}_{\mathrm{temp}}$. We say that the tower Γ_n satisfies the property of *limit multiplicity* in this case.

The limit multiplicity property is known in the case of compact quotients (also when G is a finite product of groups over local fields) by the work of DeGeorge-Wallach, Delorme and Sauvageot [13, 14, 16, 49].[3] In the non-compact case, a special case, namely the limit multiplicity for a discrete-series representation, was proved for congruence subgroups by Rohls-Speh and Savin [43, 50]. However, the complete statement about limit multiplicities was only considered in rank one cases up to now [15].

Let us explain how to extend this to congruence subgroups of GL_n. We first go back to the compact case. In that case the trace formula takes the simple form

$$\mathrm{tr}\, R_{\Gamma_n\backslash G}(f) = \mathrm{vol}(\Gamma_n\backslash G) f(1)$$

for $f \in C_c^\infty(G)$ and $n \gg 1$ (depending on the support of f). In particular,

$$\frac{\mu_n(\hat{f})}{\mathrm{vol}(\Gamma_n)} \to \mu_{\mathrm{pl}}(\hat{f})$$

where $\hat{f}(\pi) = \mathrm{tr}\,\pi(f)$. Since $\{\hat{f}\big|_{\hat{G}_{\mathrm{temp}}} : f \in C_c^\infty(G)\}$ comprise a rich space of functions on $\hat{G}_{\mathrm{temp}}$ this ultimately implies the convergence of measures on the tempered spectrum.

In the non-compact case there are additional terms, which come from the unipotent conjugacy classes on the geometric side and the continuous spectrum on the spectral side. The main problem is to show that the contribution from the continuous spectrum is negligible. To that end we use the expression (3), together with the known analytic properties of Rankin-Selberg L-functions (which

[3] Technically, a somewhat weaker result is stated in [49] but the stronger result follows from the method.

form the global constants of the intertwining operators for GL_n) to reduce to the following local problem. Given a maximal parabolic subgroup $P = MU$ of G over a p-adic field and a unitary representation σ of M it is known that the matrix coefficients $(\mathcal{F}_{\overline{P}|P}(\sigma, s)\varphi_1, \varphi_2)$ of the intertwining operators are a rational functions is q^{-s} with denominator of bounded degree [52]. The problem is to estimate the degree of the *numerator* of the matrix coefficients when φ_1, φ_2 are K-fixed, in terms of the level K. One can reduce to the case where σ is supercuspidal. A geometric argument further reduces this question to bounding the support of matrix coefficients of σ in terms of the level of the vectors. This question, in turn, is naturally answered in terms of representing σ as an induced representation from a representation on a compact subgroup modulo the center [11].

2. Periods of Automorphic Forms

A general question in automorphic forms is to study period integrals[4]

$$\int_{H(F)\backslash H(\mathbb{A})} \varphi(h)\, dh \tag{4}$$

where φ is an automorphic forms on $G(F)\backslash G(\mathbb{A})$ and H is a closed subgroup of G defined over F. In general such a period has to be regularized, but it converges in the cuspidal case.

Such periods appear in many contexts and among other things, provide a link to cohomology of locally symmetric spaces.

We say that a cuspidal representation π of $G(\mathbb{A})$ is *distinguished* by H if there exists a form φ in the space of π such that (4) does not vanish. This notion is of interest if either H is "large" or if π is "small". A particularly compelling context is the case where H is the stabilizer of a generic point of a spherical variety of G. (In general, several forms of H are needed to be considered simultaneously. Also, a frequent variant of (4) is integrating against a character of $H(\mathbb{A})$ which is trivial on $H(F)$.)

Many period integrals show up in the theory of Rankin-Selberg integrals.[5] The simplest example is the Hecke integral $\int_{F^*\backslash \mathbb{I}_F} \varphi(\left(\begin{smallmatrix} t & 0 \\ 0 & 1 \end{smallmatrix}\right))\, dt$ for a cuspidal representation π on GL_2, which by Fourier expansion unfolds to

$$L^S\left(\frac{1}{2}, \pi\right) \int_{F_S^*} \mathcal{W}^\psi(\varphi)(\left(\begin{smallmatrix} t & 0 \\ 0 & 1 \end{smallmatrix}\right))\, dt \tag{5}$$

[4] We will assume that $Z(G)$ is anisotropic for this discussion

[5] We use this terminology to mean integrals which give rise to L-functions, whether or not they involve Eisenstein series.

where for a non-trivial character ψ of $F\backslash\mathbb{A}$ we set

$$\mathcal{W}^{\psi}(\varphi)(g) = \int_{F\backslash\mathbb{A}_F} \varphi(\left(\begin{smallmatrix} 1 & x \\ 0 & 1 \end{smallmatrix}\right) g)\overline{\psi(x)}\, dx.$$

Formally, we can think of the right-hand side of (5) as the non-convergent integral $\int_{\mathbb{I}_F} \mathcal{W}(\varphi)(\left(\begin{smallmatrix} t & 0 \\ 0 & 1 \end{smallmatrix}\right))\, dt$. The general unfolding, when applicable, *formally* expresses (4) as a non-convergent integral

$$\int_{H(\mathbb{A})\cap N(\mathbb{A})\backslash H(\mathbb{A})} P\varphi(h)\, dh$$

where $P\varphi(h) = \int_{N(F)\backslash N(\mathbb{A})} \varphi(nh)\chi^{-1}(n)\, dn$ for a suitable subgroup N and a character χ of $N(\mathbb{A})$ trivial on $N(F)$. (Typically, $P\varphi$ is the Whittaker functional.) This results in an equality of the type

$$\int_{H(F)\backslash H(\mathbb{A})} \varphi(h)\, dh = L^S(s_0,\pi) \int_{(H\cap N)(F_S)\backslash H(F_S)} P\varphi(h)\, dh \qquad (6)$$

(with an appropriate regularization of the right-hand side if necessary) for a suitable special value of an L-function, depending on the setup.

It turns out however that such a relation sometimes holds even outside the context of Rankin-Selberg integrals. For instance, let D be an inner form of PGL_2, T a torus in D and let $N = D \times T$ containing T embedded diagonally. A well-known result of Waldspurger provides an identity of the form (6) for $G = N \times N$, $\pi = \sigma \otimes \chi \otimes \tilde{\sigma} \otimes \chi^{-1}$, $H = T \times T$ and N diagonally embedded in G [57]. More generally, a recent conjecture (and several important special cases) of Ichino-Ikeda extend this formalism (with precise constants) to the Gross-Prasad setup where O is either an orthogonal or a unitary group, O' is the stabilizer in O of an anisotropic vector, $N = O \times O'$ containing O' diagonally embedded, $G = N \times N$ containing N diagonally embedded, $\pi = \pi_1 \otimes \pi_2 \otimes \tilde{\pi}_1 \otimes \tilde{\pi}_2$, and $H = O' \times O'$ [28].

The formalism above, inasmuch as it can be made rigorous, provides a beautiful local-to-global principle for (4) as well as a common roof for many existing computations (and conjectures) in the theory of automorphic forms. Alternatively, we can think of this principle as expressing (4) in terms of a different period with respect to N, which is sometimes more amenable for computation (or at least provides a different model for the representation).

The period (4) defines an $H(\mathbb{A})$-invariant functional on π, and hence an $H(F_v)$-invariant functional on any local component of π. This leads to the study, in the local setup, of the space $\mathrm{Hom}_H(\pi,\mathbb{C})$ of $H(F)$-invariant functionals on π where now π is an irreducible representation of $G(F)$. In particular, when is it non-zero? and in this case, is it necessarily one-dimensional? It is only when the answer is positive that we can expect a relation such as (6) to hold without modification. These questions, which are interesting in their own right, had been studied by Gelfand, Kazhdan and Bernstein [23, 10]. Recently, there

has been a tremendous progress on these questions by Aizenbud-Gourevitch and others, and important longstanding problems were resolved [3, 2]. In turn, these results were used in the spectacular solution due to Waldspurger of the local Gross-Prasad conjectures [56].

The harmonic analysis on G/H has received much attention over the years, especially in the case of symmetric spaces. We refer the reader to [17] and the references cited therein. We also mention exciting recent work and conjectures by Sakellaridis-Venkatesh dealing with local and global aspects of periods in the context of spherical varieties [46].

Next we consider a case where there is no local uniqueness for H-invariant functionals. Remarkably, the global periods will nevertheless factorize into local invariant functionals! [31].

2.1. Unitary periods. From now on we consider a quadratic separable extension E/F with $\mathrm{Gal}(E/F)=\{1,\tau\}$ and the group $G=\mathrm{GL}_n(E)$ acting on the space X of non-degenerate hermitian forms of rank n. For any $x\in X$ the stabilizer G^x is a unitary group. In the finite field case G acts transitively on X and for an irreducible representation π of G, $\mathrm{Hom}_{G^x}(\pi,\mathbb{C})\neq 0$ if and only if $\pi^\tau\simeq\pi$, or equivalently π is obtained as a *base change* in the sense of [54] from $G'=\mathrm{GL}_n(F)$, and in this case $\mathrm{Hom}_{G^x}(\pi,\mathbb{C})$ is one-dimensional [24].

Consider the case where F is a local field of characteristic 0. There are finitely many orbits of G on X. (Exactly two in the non-archimedean case.) Let ω be the quadratic character of F^* corresponding to E under class field theory. We denote by bc the base change map from the irreducible representations of G' to those of G [9]. The map is defined for any cyclic extension and is characterized (for tempered representations) by certain character identities. The image of bc is the Galois invariant representations.

We will describe an analogue of these character identities where the characters are replaced by spherical characters which are distributions occurring in the relative trace formula and satisfying certain invariance properties. On the G'-side they are left and right equivariant with respect a non-degenerate character ψ' on group of upper unitriangular matrices $U_0(F)$. On the G-side they are right invariant under a unitary group and left $(U_0(E),\psi)$-equivariant where $\psi(u)=\psi'(uu^\tau)$.

The base change is a local counterpart (and consequence) of a global correspondence and the latter is proved by comparing a trace formula for $G\rtimes\mathrm{Gal}(E/F)$ with the trace formula for G'. In analogy, the new character identities are also obtained by global means using a comparison of the relative trace formula. The comparison in this context was introduced by Jacquet. To describe it let E/F be a quadratic extension of number fields and consider on the one hand

$$\int_{U_0(E)\backslash U_0(\mathbb{A}_E)} K_\Phi(u)\psi(u)\,du$$

where $\Phi \in \mathcal{S}(X(\mathbb{A}))$, U_0 is the group of upper unitriangular matrices, ψ is a non-degenerate character of U_0 and

$$K_\Phi(g) = \sum_{x \in X(F)} \Phi({}^t g^\tau x g) \quad g \in \mathrm{GL}_n(\mathbb{A}_E).$$

It is compared with a Kuznetsov trace formula

$$\iint_{(U_0(F)\backslash U_0(\mathbb{A}))^2} K_f(u_1, u_2)\psi'(u_1^{-1}u_2)\, du_1\, du_2$$

for appropriate $f \in \mathcal{S}(G'(\mathbb{A}))$ which matches Φ in the sense that certain orbital integrals are equal. The geometric comparison, the existence of matching functions, and the relevant fundamental Lemma are worked out in [33, 32, 1]. The spectral expansion into absolutely convergent terms is carried out in [37]. The contribution of the continuous spectrum involves certain regularized periods which are studied in [38].

We will now describe the results of [20] extending previous results of Jacquet. We return to the local setup and consider for any representation π of G the space

$$\mathfrak{U}(\pi) = \mathrm{Hom}_G(\mathcal{S}(X), \pi^*) \simeq \mathrm{Equiv}_G(X, \pi^*) \simeq \oplus_{x \in X/G} \mathrm{Hom}_{G^x}(\pi, \mathbb{C})$$

where Equiv_G denotes the space of G-equivariant maps.

Theorem 2.1. *Let π' be a unitary generic irreducible representation of $\mathrm{GL}_n(F)$ with Whittaker functional $\mathcal{W}'$ and let $\pi = \mathrm{bc}(\pi')$. Then there exists $\alpha^{\pi'} \in \mathfrak{U}(\pi)$ such that*

$$\sum \alpha^{\pi'}(\Phi)(v)\mathcal{W}(v) = \sum \mathcal{W}'(\pi'(f)v')\mathcal{W}'(v')$$

for matching $\Phi \leftrightarrow f$ where the sums are over orthonormal bases of π and π' respectively. Moreover, let $\alpha_x^{\pi'}$, $x \in X$ be the corresponding G^x-invariant functional of π. Then in the p-adic case $\alpha_x^{\pi'} \equiv 0$ if and only if $\pi' \simeq \pi' \otimes \omega$ and G^x is not quasi-split.

In fact, it is possible to define $\alpha^{\pi'} \in \mathfrak{U}(\pi)$ for any irreducible generic representation π' (not necessarily unitary). Note that the split analogue of $\alpha^{\pi'}$ is the pairing $\pi \otimes \pi^\iota \to \mathbb{C}$ where $\iota(g) = {}^t g^{-1}$.

In the archimedean case we can write π' as induced from the quasi-characters $\chi_1, \ldots, \chi_k$ of $\mathbb{R}^*$ and the essentially square-integrable representations $\sigma_1, \ldots, \sigma_l$ of $\mathrm{GL}_2(\mathbb{R})$ with $n = k + 2l$. For any quasi-character χ of $\mathbb{R}^*$ let $m_{\pi'}(\chi) = \#\{i : \chi_i = \chi\}$ and define the integer $\mathfrak{d}_{\pi'} = l + \sum_{s \in \mathbb{C}} \min(m_{\pi'}(|\cdot|^s), m_{\pi'}(|\cdot|^s \mathrm{sgn})) \leq n/2$.

Conjecture 2.2. *In the archimedean case $\alpha_x^{\pi'} \not\equiv 0$ if and only if $\mathfrak{d}_{\pi'} \leq \mathrm{rk}(G^x)$ where $\mathrm{rk}\, U(p,q) = \min(p,q)$.*

Consider the tempered representations $\pi_{k,l}$, on $\mathrm{GL}_n(\mathbb{R})$, $n = k + l$ induced from the character $(t_1, \dots, t_n) \mapsto \mathrm{sgn}(t_1 \dots t_l)$ of the Borel subgroup. They comprise the fiber under bc of the representation of $\mathrm{GL}_n(\mathbb{C})$ induced from the trivial character of the Borel. For any p, q with $n = p + q$ let $\Omega_{p,q}$ be the open subset $N_0(\mathbb{R})w_0 D_{p,q} N_0(\mathbb{R})$ of $\mathrm{GL}_n(\mathbb{R})$ where $D_{p,q} = \{\mathrm{diag}(t_1, \dots, t_n) \in T_0(\mathbb{R}) : \sum_{i=1}^n \mathrm{sgn}\, t_i = p - q\}$ and w_0 is the permutation matrix with ones on the secondary diagonal. Conjecture 2.2 would follow from the following concrete conjecture.

Conjecture 2.3. *The restriction of the distribution* $f \mapsto \sum \mathcal{W}(\pi_{k,l}(f)v')\mathcal{W}(v')$ *to* $\Omega_{p,q}$ *is non-zero if and only if* $\min(k, l) \le \min(p, q)$.

The "only if" part of conjectures 2.2 and 2.3 is known. The "if" part is known in the case of quasi-split unitary group [30].

The global result for unitary periods is the following.

Theorem 2.4. *Let* x *be a hermitian form. A cuspidal representation* π *of* $G(\mathbb{A})$ *is distinguished by* G^x *if and only if* π *is the base change of a cuspidal representation* π' *of* $\mathrm{GL}_n(\mathbb{A})$ *and* $\alpha_x^{\pi'_v} \not\equiv 0$ *for all inert places* v *of* F. *In particular, any Galois invariant* π *is distinguished by the quasi-split unitary group. Moreover, for a suitable normalization of Haar measures we have*

$$\int_{G^x(F)\backslash G^x(\mathbb{A})} \varphi(h)\, dh = 2L^S(1, \pi' \otimes \tilde{\pi}' \otimes \omega) \prod_{v \in S} \alpha_x^{\pi'_v}(\varphi_v)$$

where $\varphi = \otimes \varphi_v$ *is a factorizable vector in* $\pi = \otimes \pi_v$ *and* S *is a sufficiently large set of places of* F.

For an application towards Sarnak's conjecture on L^∞-norm of automorphic form see [36].

We go back to the local setup and analyze the space $\mathfrak{U}(\pi)$ more carefully. We denote by $\pi_1 \times \pi_2$ the parabolic induction from $\pi_1 \otimes \pi_2$. Suppose that F is p-adic. We first reduce to the case where π is *pure*, i.e. a subquotient of $\sigma_1 \times \dots \times \sigma_k$ where the σ_i's are unramified twists of a single Galois invariant supercuspidal representation. More precisely, we have

Theorem 2.5. *Let* π *be an irreducible representation of* G. *Then*

1. $\mathfrak{U}(\pi)$ *is finite-dimensional.*
2. *If* $\mathfrak{U}(\pi) \neq 0$ *then* π *is Galois invariant.*
3. *Let* σ *be an irreducible subquotient of* $\sigma_1 \times \dots \times \sigma_k$ *where* σ_i *are essentially supercuspidal and not Galois invariant. Then* $\mathfrak{U}(\sigma \times \sigma^\tau)$ *is one-dimensional and* $\mathfrak{U}(\sigma \times \sigma^\tau \times \pi) \simeq \mathfrak{U}(\pi)$.
4. *Suppose that* π_1, π_2 *are Galois invariant with disjoint supercuspidal support. Then* $\mathfrak{U}(\pi_1 \times \pi_2) \simeq \mathfrak{U}(\pi_1) \otimes \mathfrak{U}(\pi_2)$.

An irreducible representation π is called *imprimitive* if it is the Langlands quotient of $\delta_1 \times \cdots \times \delta_k$ where δ_i are essentially square-integrable and $\delta_i \times \delta_{i+1}$ is reducible for all i. This is an important class of representations containing the square-integrable ones and the general Speh representations. Note that every Galois invariant imprimitive representation is pure.

Theorem 2.6. *For any imprimitive π and $x \in X$ we have* $\dim \mathrm{Hom}_{G^x}(\pi, \mathbb{C}) \leq 1$.

We expect that $\dim \mathrm{Hom}_{G^x}(\pi, \mathbb{C}) = 1$ if π is imprimitive and Galois invariant. This is the case if the kernel of the quotient map $\delta_1 \times \cdots \times \delta_k \to \pi$ (the longest intertwining operator) is spanned by the kernels of the rank one intertwining operators – a property which is likely to hold in general for imprimitive representations.

In a special case, we can determine multiplicities completely.

Theorem 2.7. *Suppose that $\pi = \delta_1 \times \cdots \times \delta_k$ where $\delta_1, \ldots, \delta_k$ are essentially square-integrable, Galois invariant and distinct. Then* $\dim \mathfrak{U}(\pi) = 2^k = 2 \dim \mathrm{Hom}_{G^x}(\pi, \mathbb{C})$ *for all $x \in X$. Moreover, $\{\alpha^{\pi'} : \mathrm{bc}(\pi') = \pi\}$ is a basis for $\mathfrak{U}(\pi)$.*

It is also true in general that $\dim \mathfrak{U}(\pi_1 \times \pi_2) \geq \dim \mathfrak{U}(\pi_1) \dim \mathfrak{U}(\pi_2)$. We do not know whether an equality always holds if π_1, π_2 are Galois invariant.

In the archimedean case we have

Theorem 2.8 (Aizenbud-Lapid). *Let π be the Langlands quotient of $\chi_1 \times \cdots \times \chi_n$, where $\chi_1, \ldots \chi_n \in \widehat{\mathbb{C}^*}$. Suppose that π is distinguished by $U(p, q)$. Then $\pi^\tau \simeq \pi$ (i.e.* $\mathrm{Gal}(\mathbb{C}/\mathbb{R})$ *stabilizes the multiset $\{\chi_1, \ldots, \chi_n\}$) and the number of* $\mathrm{Gal}(\mathbb{C}/\mathbb{R})$*-orbits of size two does not exceed* $\min(p, q)$.

We expect that the converse is also true. This is known in the generic case, i.e., when $\chi_1 \times \cdots \times \chi_n$ is irreducible.

References

[1] Avraham Aizenbud and Dmitry Gourevitch. Smooth transfer of Kloosterman integrals (the archimedean case). preprint.

[2] Avraham Aizenbud and Dmitry Gourevitch. Generalized Harish-Chandra descent, Gelfand pairs, and an Archimedean analog of Jacquet-Rallis's theorem. *Duke Math. J.*, 149(3):509–567, 2009. With an appendix by the authors and Eitan Sayag.

[3] Avraham Aizenbud, Dmitry Gourevitch, Rallis Stephen, and Gérard Schiffmann. Multiplicity one theorems. *Ann. of Math. (2)*, to appear.

[4] James Arthur. On a family of distributions obtained from Eisenstein series. II. Explicit formulas. *Amer. J. Math.*, 104(6):1289–1336, 1982.

[5] James Arthur. A measure on the unipotent variety. *Canad. J. Math.*, 37(6):1237–1274, 1985.

[6] James Arthur. On a family of distributions obtained from orbits. *Canad. J. Math.*, 38(1):179–214, 1986.

[7] James Arthur. The local behaviour of weighted orbital integrals. *Duke Math. J.*, 56(2):223–293, 1988.

[8] James Arthur. An introduction to the trace formula. In *Harmonic analysis, the trace formula, and Shimura varieties*, volume 4 of *Clay Math. Proc.*, pages 1–263. Amer. Math. Soc., Providence, RI, 2005.

[9] James Arthur and Laurent Clozel. *Simple algebras, base change, and the advanced theory of the trace formula*, volume 120 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 1989.

[10] Joseph N. Bernstein. P-invariant distributions on $\mathrm{GL}(N)$ and the classification of unitary representations of $\mathrm{GL}(N)$ (non-Archimedean case). In *Lie group representations, II (College Park, Md., 1982/1983)*, volume 1041 of *Lecture Notes in Math.*, pages 50–102. Springer, Berlin, 1984.

[11] Colin J. Bushnell and Philip C. Kutzko. *The admissible dual of* $\mathrm{GL}(N)$ *via compact open subgroups*, volume 129 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 1993.

[12] Laurent Clozel and Patrick Delorme. Le théorème de Paley-Wiener invariant pour les groupes de Lie réductifs. II. *Ann. Sci. École Norm. Sup. (4)*, 23(2):193–228, 1990.

[13] David L. de George and Nolan R. Wallach. Limit formulas for multiplicities in $L^2(\Gamma\backslash G)$. *Ann. of Math. (2)*, 107(1):133–150, 1978.

[14] David L. DeGeorge and Nolan R. Wallach. Limit formulas for multiplicities in $L^2(\Gamma\backslash G)$. II. The tempered spectrum. *Ann. of Math. (2)*, 109(3):477–495, 1979.

[15] Anton Deitmar and Werner Hoffmann. On limit multiplicities for spaces of automorphic forms. *Canad. J. Math.*, 51(5):952–976, 1999.

[16] Patrick Delorme. Formules limites et formules asymptotiques pour les multiplicités dans $L^2(G/\Gamma)$. *Duke Math. J.*, 53(3):691–731, 1986.

[17] Patrick Delorme. Harmonic analysis on real reductive symmetric spaces. In *Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002)*, pages 545–554. Higher Ed. Press, Beijing, 2002.

[18] Harold Donnelly. On the cuspidal spectrum for finite volume symmetric spaces. *J. Differential Geom.*, 17(2):239–253, 1982.

[19] J. J. Duistermaat, J. A. C. Kolk, and V. S. Varadarajan. Spectra of compact locally symmetric manifolds of negative curvature. *Invent. Math.*, 52(1):27–93, 1979.

[20] Brooke Feigon, Erez Lapid, and Omer Offen. On representations distinguished by unitary groups. *preprint*.

[21] Tobias Finis and Erez Lapid. On the continuity of Arthur's trace formula: the elliptic terms. *Compos. Math.*, to appear.

[22] Tobias Finis, Erez M. Lapid, and Werner Müller. The spectral side of Arthur's trace formula. *Proc. Natl. Acad. Sci. USA*, 106(37):15563–15566, 2009.

[23] I. M. Gel′fand and D. A. Kajdan. Representations of the group $\mathrm{GL}(n, K)$ where K is a local field. In *Lie groups and their representations (Proc. Summer School, Bolyai János Math. Soc., Budapest, 1971)*, pages 95–118. Halsted, New York, 1975.

[24] Roderick Gow. Two multiplicity-free permutation representations of the general linear group $\mathrm{GL}(n, q^2)$. *Math. Z.*, 188(1):45–54, 1984.

[25] Sigurdur Helgason. *Geometric analysis on symmetric spaces*, volume 39 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1994.

[26] Werner Hoffmann. The nonsemisimple term in the trace formula for rank one lattices. *J. Reine Angew. Math.*, 379:1–21, 1987.

[27] Lars Hörmander. The spectral function of an elliptic operator. *Acta Math.*, 121:193–218, 1968.

[28] Atsushi Ichino and Tamotsu Ikeda. On the periods of automorphic forms on special orthogonal groups and the Gross-Prasad conjecture. *Geom. Funct. Anal.*, 19(5):1378–1425, 2010.

[29] Henryk Iwaniec, Wenzhi Luo, and Peter Sarnak. Low lying zeros of families of L-functions. *Inst. Hautes Études Sci. Publ. Math.*, (91):55–131 (2001), 2000.

[30] Hervé Jacquet. Distinction by a quasi-split unitary group. *Israel J. Math.*, to appear.

[31] Hervé Jacquet. Factorization of period integrals. *J. Number Theory*, 87(1):109–143, 2001.

[32] Hervé Jacquet. Smooth transfer of Kloosterman integrals. *Duke Math. J.*, 120(1):121–152, 2003.

[33] Hervé Jacquet. Kloosterman identities over a quadratic extension. II. *Ann. Sci. École Norm. Sup. (4)*, 38(4):609–669, 2005.

[34] Robert P. Langlands. *On the functional equations satisfied by Eisenstein series*. Springer-Verlag, Berlin, 1976. Lecture Notes in Mathematics, Vol. 544.

[35] Erez Lapid and Werner Müller. Spectral asymptotics for arithmetic quotients of $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$. *Duke Math. J.*, 149(1):117–155, 2009.

[36] Erez Lapid and Omer Offen. Compact unitary periods. *Compos. Math.*, 143(2):323–338, 2007.

[37] Erez M. Lapid. On the fine spectral expansion of Jacquet's relative trace formula. *J. Inst. Math. Jussieu*, 5(2):263–308, 2006.

[38] Erez M. Lapid and Jonathan D. Rogawski. Periods of Eisenstein series: the Galois case. *Duke Math. J.*, 120(1):153–226, 2003.

[39] Elon Lindenstrauss and Akshay Venkatesh. Existence and Weyl's law for spherical cusp forms. *Geom. Funct. Anal.*, 17(1):220–251, 2007.

[40] Wenzhi Luo. Nonvanishing of L-values and the Weyl law. *Ann. of Math. (2)*, 154(2):477–502, 2001.

[41] W. Müller. On the spectral side of the Arthur trace formula. *Geom. Funct. Anal.*, 12(4):669–722, 2002.

[42] Werner Müller. The trace class conjecture in the theory of automorphic forms. *Ann. of Math. (2)*, 130(3):473–529, 1989.

[43] Jürgen Rohlfs and Birgit Speh. On limit multiplicities of representations with cohomology in the cuspidal spectrum. *Duke Math. J.*, 55(1):199–211, 1987.

[44] Hiroshi Saito. Automorphic forms and algebraic extensions of number fields. *Proc. Japan Acad.*, 51(4):229–233, 1975.

[45] Hiroshi Saito. Convergence of the zeta functions of prehomogeneous vector spaces. *Nagoya Math. J.*, 170:1–31, 2003.

[46] Yiannis Sakellaridis and Akshay Venkatesh. Periods and harmonic anaysis on spherical varieties. *preprint*.

[47] Peter Sarnak. Spectra of hyperbolic surfaces. *Bull. Amer. Math. Soc. (N.S.)*, 40(4):441–478 (electronic), 2003.

[48] Mikio Sato and Takuro Shintani. On zeta functions associated with prehomogeneous vector spaces. *Ann. of Math. (2)*, 100:131–170, 1974.

[49] François Sauvageot. Principe de densité pour les groupes réductifs. *Compositio Math.*, 108(2):151–184, 1997.

[50] Gordan Savin. Limit multiplicities of cusp forms. *Invent. Math.*, 95(1):149–159, 1989.

[51] A. Selberg. Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. *J. Indian Math. Soc. (N.S.)*, 20:47–87, 1956.

[52] Freydoon Shahidi. On certain L-functions. *Amer. J. Math.*, 103(2):297–355, 1981.

[53] Hideo Shimizu. On traces of Hecke operators. *J. Fac. Sci. Univ. Tokyo Sect. I*, 10:1–19 (1963), 1963.

[54] Takuro Shintani. Two remarks on irreducible characters of finite general linear groups. *J. Math. Soc. Japan*, 28(2):396–414, 1976.

[55] Takuro Shintani. On liftings of holomorphic cusp forms. In *Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 2*, Proc. Sympos. Pure Math., XXXIII, pages 97–110. Amer. Math. Soc., Providence, R.I., 1979.

[56] J.-L. Waldspurger. La conjecture locale de Gross-Prasad pour les représentations tempérées des groupes spéciaux orthogonaux. *preprint*.

[57] J.-L. Waldspurger. Sur les valeurs de certaines fonctions L automorphes en leur centre de symétrie. *Compositio Math.*, 54(2):173–242, 1985.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Finite W-algebras

Ivan Losev*

Abstract

A finite W-algebra is an associative algebra constructed from a semisimple Lie algebra and its nilpotent element. In this survey we review recent developments in the representation theory of W-algebras. We emphasize various interactions between W-algebras and universal enveloping algebras.

Mathematics Subject Classification (2010). Primary 16G99, 17B35; Secondary 53D20, 53D55.

Keywords. W-algebra, semisimple Lie algebra, nilpotent orbit, universal enveloping algebra, primitive ideal, Whittaker module.

1. Introduction

Our base field $\mathbb{K}$ is supposed to be algebraically closed and of characteristic zero.

A finite W-algebra is an associative algebra constructed from a pair $(\mathfrak{g}, e)$, where $\mathfrak{g}$ is a finite dimensional semisimple Lie algebra, and e is a nilpotent element of $\mathfrak{g}$. A W-algebra should be thought as a generalization of the universal enveloping algebra $U(\mathfrak{g})$. The latter can be considered as the W-algebra for the pair $(\mathfrak{g}, 0)$.

The study of W-algebras traces back to the celebrated paper [32] of Kostant. This paper essentially treats the case when the element e is principal (i.e., the adjoint orbit of e is dense in the nilpotent cone of $\mathfrak{g}$). Kostant's motivation came basically from the study of Whittaker vectors and of Whittaker models. [32] was followed by the thesis [42] of Lynch who was a student of Kostant. In [42] Kostant's results were (partially) generalized to arbitrary *even* nilpotent elements. During the 80's Whittaker models (in the sense different from

*Supported by the NSF grant DMS-0900907

Massachusetts Institute of Technology, Department of Mathematics, 77 Massachusetts Avenue, Cambridge MA 02139, USA. E-mail: ivanlosev@math.mit.edu.

Kostant's) were also considered in [46],[47], where they were applied to the study of certain primitive ideals in $U(\mathfrak{g})$.

In the 90's finite W-algebras attracted some attention from mathematical physicists, see, for example, [6],[56],[59]. One of the main motivations for their interest was a relationship between finite and *affine* W-algebras. The latter are certain vertex algebras modeling the so called W-symmetry from Conformal field theory.

In [50] Premet gave a general definition of a W-algebra. Premet's interest to the subject was motivated by the study of non-restricted representations of semisimple Lie algebras in positive characteristic. The paper [50] initiated a lot of work on different, mostly representation theoretic, aspects of W-algebras.

Apart from being of independent interest, finite W-algebras have several connections to other objects studied in Representation theory. Let us summarize these connections.

A) It seems that the most straightforward connection is to the universal enveloping algebras of semisimple Lie algebras. This connection can be informally explained as follows. According to the Orbit method, to an infinite dimensional representation of $\mathfrak{g}$ one should be able to assign a nilpotent orbit in $\mathfrak{g}^*(\cong \mathfrak{g})$. For instance, to a "nice" (e.g., irreducible) Harish-Chandra $\mathfrak{g}$-bimodule one assigns a dense orbit in its associated variety. Then there is a hope (that sometimes converts into proofs) that one can reduce the study of an infinite dimensional $\mathfrak{g}$-module to the study of a *finite dimensional* module over the W-algebra corresponding to the nilpotent orbit in interest. A relationship between the W-algebras and $U(\mathfrak{g})$ is studied, for example, in [10],[26],[37]–[40],[51]–[53],[57].

B) There is a connection between the representation theory of W-algebras (in characteristic zero) and that of semisimple Lie algebras in positive characteristic. In a sentence, any reduced enveloping algebra turns out to be Morita equivalent to an appropriate reduced W-algebra. One can relate the representation theories of W-algebras in positive and in zero characteristics. This relationship was successfully used in Premet's papers, see [49],[50],[52],[53].

C) For classical Lie algebras there is a connection between W-algebras and (twisted) Yangians. This connection was first discovered in [56] and then studied further in [7],[8],[12],[13],[55]. Also W-algebras are related to the (cyclotomic quotients of) degenerate affine Hecke algebras, [14].

D) As we mentioned above, finite W-algebras are related to their affine counterparts. This relation can be made formal. To any vertex algebra one can assign an associative algebra called the *Zhu algebra*. The importance of the Zhu algebra is that its representation theory controls much of the representation theory of the initial vertex algebra. A finite W-algebra is closely related to the Zhu algebra of the corresponding affine W-algebra. For details the reader is referred to [17].

In the present paper we are mostly interested in A). We also briefly explain C), while B) remains almost untouched and we do not discuss D) at all. Therefore we suppress the adjective "finite" while speaking about W-algebras.

Another review on W-algebras [60] by W. Wang have already appeared. Some topics not discussed (or discussed very briefly) in our survey can be found there.

This paper is organized as follows. In Section 2 we discuss topics related to a definition of a W-algebra via Hamiltonian reduction, which is essentially due to Premet. In Section 3 we explain the definition of a W-algebra based on the Deformation quantization, [37]. The next three sections describe connections between W-algebras and $U(\mathfrak{g})$. In Section 4 we discuss category equivalences between certain categories of modules for W-algebras and for $U(\mathfrak{g})$. Section 5 describes a relationship between the sets of two-sided ideals in the two algebras. This description leads to a (partial) classification of irreducible finite dimensional representations of W-algebras in terms of primitive ideals of $U(\mathfrak{g})$. Section 6 deals with one-dimensional modules over W-algebras. Finally in Section 7 we explain the connection C) above mostly for $\mathfrak{g}$ of type A.

In the beginning of each section its content is described in more detail.

Acknowledgements. First of all, I would like to thank J. Brundan, V. Ginzburg, S. Goodwin, A. Kleshchev, and A. Premet for numerous inspiring discussions on W-algebras. I also thank J. Brundan, V. Ginzburg, and A. Premet for their remarks on a preliminary version of this text.

Notation and conventions. Throughout the paper G is a connected reductive group, $\mathfrak{g}$ is its Lie algebra. We choose a nilpotent element $e \in \mathfrak{g}$ and pick $h, f \in \mathfrak{g}$ forming an $\mathfrak{sl}_2$-triple with e, i.e., $[h,e] = 2e, [h,f] = -2f, [e,f] = h$. Let $\mathbb{O}$ denote the G-orbit of e. Also we fix a G-invariant non-degenerate symmetric form $(\cdot,\cdot)$ on $\mathfrak{g}$. Using this form, we identify $\mathfrak{g}$ with $\mathfrak{g}^*$.

We write $\mathcal{U}$ for the universal enveloping algebra $U(\mathfrak{g})$ of $\mathfrak{g}$. By $\mathcal{Z}$ we denote the center of $\mathcal{U}$. This is a polynomial algebra.

Let us also list some standard notation used below.

A^{op}	the opposite algebra of an algebra A.
$\mathrm{Ann}_A(M)$	the annihilator of an A-module M.
$\mathrm{End}_A(M)$	the algebra of endomorphisms of an A-module M.
$\mathrm{End}(M)$	$:= \mathrm{End}_{\mathbb{K}}(M)$.
$\mathrm{gr}\, V$	the associated graded vector space of a filtered vector space V.
H°	the unit component of an algebraic group H.
$\mathbb{K}[X]$	the algebra of regular functions on a variety X.
$\mathbb{K}[X]^\wedge_Y$	the algebra of functions on the completion of a variety X along a subvariety Y.
T^*X	the cotangent bundle of a smooth variety X.
$\mathrm{V}(\mathcal{M})$	the associated variety of a finitely generated $\mathcal{U}$-module $\mathcal{M}$.
$\mathfrak{z}(\mathfrak{h})$	the center of a Lie algebra $\mathfrak{h}$.
$\mathfrak{z}_\mathfrak{h}(\mathfrak{f})$	the centralizer of $\mathfrak{f}$ in a Lie algebra $\mathfrak{h}$.

2. W-algebras Via Hamiltonian Reduction

In this section we discuss developments leading to and related to Premet's definition of a W-algebra given in [50]. The first such development is, of course, Kostant's work, [32], where the case of a principal nilpotent element was treated. We describe (very few of) Kostant's results in Subsection 2.1. Then in Subsection 2.2 we mention a generalization of Kostant's constructions to the case of an even nilpotent element due to Lynch, [42]. In Subsection 2.3 we provide Premet's definition in the form of a quantum Hamiltonian reduction. In Subsection 2.4 we show that the "quasiclassical limit" of a W-algebra is the algebra of functions on a *Slodowy slice* that is a transverse slice to a nilpotent orbit in $\mathfrak{g}$ introduced in [58]. In Subsection 2.5 we mention several ramifications of Premet's definition and in Subsection 2.6 discuss some properties of W-algebras that can be proved using this definition.

2.1. Kostant's results: the case of a principal nilpotent element. In this subsection we will explain (some of) Kostant's results, Section 2 of [32].

Suppose the nilpotent element e is principal. Set

$$\mathfrak{g}(i) := \{\xi \in \mathfrak{g} | [h, \xi] = i\xi\}, \mathfrak{p} := \bigoplus_{i \geqslant 0} \mathfrak{g}(i), \mathfrak{m} := \bigoplus_{i<0} \mathfrak{g}(i), \chi := (e, \cdot). \tag{2.1}$$

Let us describe $\mathfrak{p}, \mathfrak{m}, e, \chi$ in more conventional terms. Let $\mathfrak{h} \subset \mathfrak{g}$ be a Cartan subalgebra of $\mathfrak{g}$, $\Delta \subset \mathfrak{h}^*$ the corresponding root system, and Π a system of simple roots in Δ. Further, for $\alpha \in \Delta$ let e_α denote a corresponding weight vector in $\mathfrak{g}$. Finally, let $\rho^\vee$ denote half the sum of all positive coroots (=the sum of all fundamental co-weights). Replacing (e, h, f) with a G-conjugate triple, we may assume that $h = 2\rho^\vee$ and $e = \sum_{\alpha \in \Pi} e_\alpha$. So $\mathfrak{p}$ becomes the positive Borel subalgebra $\mathfrak{b} \subset \mathfrak{g}$, $\mathfrak{m}$ becomes the negative maximal nilpotent subalgebra $\mathfrak{n}_-$, while χ is a non-degenerate character of $\mathfrak{m}$.

Define the shift $\mathfrak{m}_\chi := \{\xi - \langle \chi, \xi \rangle, \xi \in \mathfrak{m}\}$ of $\mathfrak{m}$. Then, thanks to the PBW theorem, we get

$$\mathcal{U} = U(\mathfrak{p}) \oplus \mathcal{U}\mathfrak{m}_\chi. \tag{2.2}$$

Using this decomposition, one can define an action of $\mathfrak{m}$ on $U(\mathfrak{p})$. Namely, identify $U(\mathfrak{p})$ with the quotient $\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi$ using (2.2). The adjoint action of $\mathfrak{m}$ on $\mathcal{U}$ descends to $\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi$. Using the identification $U(\mathfrak{p}) \cong \mathcal{U}/\mathcal{U}\mathfrak{m}_\chi$, we get an $\mathfrak{m}$-action on $U(\mathfrak{p})$.

By definition, a W-algebra $U(\mathfrak{g}, e)$ is the invariant subalgebra $U(\mathfrak{p})^{\mathfrak{m}}$. In other words, $U(\mathfrak{g}, e)$ is the *quantum Hamiltonian reduction*

$$(\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi)^{\operatorname{ad}\mathfrak{m}} := \{a + \mathcal{U}\mathfrak{m}_\chi : [\xi, a] \in \mathcal{U}\mathfrak{m}_\chi, \forall \xi \in \mathfrak{m}\}.$$

The multiplication on the last space is defined by $(a + \mathcal{U}\mathfrak{m}_\chi)(b + \mathcal{U}\mathfrak{m}_\chi) = ab + \mathcal{U}\mathfrak{m}_\chi$.

It turns out that $U(\mathfrak{g},e)$ is naturally isomorphic to the center $\mathcal{Z}$ of $\mathcal{U}$. Namely, the inclusion $\mathcal{Z}\hookrightarrow\mathcal{U}$ gives rise to a natural map $\mathcal{Z}\to\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi$. Its image clearly consists of $\mathfrak{m}$-invariants. So we get a homomorphism $\mathcal{Z}\to U(\mathfrak{g},e)$. By Theorem 2.4.1 in [32], this homomorphism is an isomorphism. In particular, we get an embedding of $\mathcal{Z}$ into $U(\mathfrak{p})$. This embedding is of importance in the quantization of Toda systems, see [33].

2.2. Generalization: the case of even e. Recall that e is called *even* if all eigenvalues of $\operatorname{ad}h$ on $\mathfrak{g}$ are even. Define $\mathfrak{g}(i),\mathfrak{p},\mathfrak{m},\chi$ by (2.1). It is clear that $\mathfrak{p}$ is a parabolic subalgebra of $\mathfrak{g}$ and $\mathfrak{m}$ is the nilpotent radical of the opposite parabolic. In [42] Lynch generalized Kostant's definition and introduced an algebra $U(\mathfrak{g},e):=U(\mathfrak{p})^{\mathfrak{m}}=(\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi)^{\operatorname{ad}\mathfrak{m}}$.

There is an embedding $U(\mathfrak{g},e)\hookrightarrow U(\mathfrak{g}(0))$ sometimes called the *generalized Miura transform*. It is obtained by restricting the natural projection $U(\mathfrak{p})\twoheadrightarrow U(\mathfrak{g}(0))$ to $U(\mathfrak{g},e)\subset U(\mathfrak{p})$. The restriction is injective by [42], Corollary 2.3.2.

2.3. Definition of $U(\mathfrak{g},e)$: the general case. Now let $e\in\mathfrak{g}$ be an arbitrary (nonzero) nilpotent element. Let the decomposition $\mathfrak{g}=\bigoplus_{i\in\mathbb{Z}}\mathfrak{g}(i)$ and the element $\chi\in\mathfrak{g}^*$ be given by (2.1). Following Premet, [50], we still can define a W-algebra $U(\mathfrak{g},e)$ as the quantum Hamiltonian reduction $(\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi)^{\operatorname{ad}\mathfrak{m}}$ provided we can find a suitable analog of the subalgebra $\mathfrak{m}\subset\mathfrak{g}$ considered in the previous subsection.

A subalgebra $\mathfrak{m}$ we need is constructed as follows. Consider a skew-symmetric form ω_χ on $\mathfrak{g}$ given by $\omega_\chi(\xi,\eta)=\langle\chi,[\xi,\eta]\rangle$. It follows easily from the representation theory of $\mathfrak{sl}_2$ that the restriction of ω_χ to the subspace $\mathfrak{g}(-1)$ is non-degenerate. Pick a lagrangian subspace $l\subset\mathfrak{g}(-1)$ and set $\mathfrak{m}:=l\oplus\bigoplus_{i\leqslant-2}\mathfrak{g}(i)$.

It is clear that $\mathfrak{m}$ is a subalgebra in $\mathfrak{g}$ consisting of nilpotent elements. Also since ω_χ vanishes on l, we see that $\langle\chi,[\mathfrak{m},\mathfrak{m}]\rangle=0$. So χ is indeed a character of $\mathfrak{m}$.

We set $U(\mathfrak{g},e):=(\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi)^{\operatorname{ad}\mathfrak{m}}$. The reader should notice that, a priory, this definition is ambiguous: $\mathfrak{m}$ and hence $U(\mathfrak{g},e)$ depend on the choice of l. However, we will see in Subsection 2.5 that two W-algebras constructed using different choices of l are canonically isomorphic.

We finish the subsection with a few historical remarks. Direct analogs of $\mathfrak{m}$ and of its shift $\mathfrak{m}_\chi$ in the setting of finite Chevalley groups first appeared in [30]. Then Moeglin used $\mathfrak{m}_\chi$ to define "Whittaker models" for primitive ideals of $U(\mathfrak{g})$, [46],[47]. We will describe her results in more detail in Section 6. Later the subalgebra $\mathfrak{m}$ played an important role in Premet's proof of the Kac-Weisfeller conjecture on the dimension of a non-restricted representation of a semisimple Lie algebra in positive characteristic, see [49].

2.4. Classical counterpart: the Slodowy slice. The algebra $U(\mathfrak{g},e)$ has an interesting filtration, called the *Kazhdan* filtration.

To define it we first introduce a new filtration on $\mathcal{U}$. Recall that the algebra $\mathcal{U}$ has the standard, PBW filtration: the subspace $\mathrm{F}_i^{st}\mathcal{U}$ of elements of degree $\leqslant i$, by definition, is spanned by all monomials $\xi_1\dots\xi_j, j\leqslant i, \xi_1,\dots,\xi_j\in\mathfrak{g}$. For $j\in\mathbb{Z}$ set $\mathcal{U}(j):=\{u\in\mathcal{U}|[h,u]=ju\}$. Define the *Kazhdan filtration* $\mathrm{K}_i\mathcal{U}$ on $\mathcal{U}$ by $\mathrm{K}_i\mathcal{U}:=\sum_{2j+k\leqslant i}\mathrm{F}_j^{st}\mathcal{U}\cap\mathcal{U}(k)$. We remark that the associated graded algebra of $\mathcal{U}$ with respect to the Kazhdan filtration is still naturally isomorphic to the symmetric algebra $S(\mathfrak{g})$.

Being a subquotient of $\mathcal{U}$, the algebra $U(\mathfrak{g},e)$ has a Kazhdan filtration $\mathrm{K}_i U(\mathfrak{g},e)$ inherited from $\mathcal{U}$. We remark that $\mathrm{K}_0\mathcal{U}\subset\mathbb{K}+\mathcal{U}\mathfrak{m}_\chi$ so the Kazhdan filtration on $U(\mathfrak{g},e)$ is positive in the sense that $\mathrm{K}_0 U(\mathfrak{g},e)=\mathbb{K}$.

It turns out that the associated graded algebra $\operatorname{gr}U(\mathfrak{g},e)$ of $U(\mathfrak{g},e)$ is naturally isomorphic to the algebra of functions on the *Slodowy slice* $S:=e+\ker\operatorname{ad}(f)$, [58] (in the case when e is principal S appeared in [31]). It follows from the representation theory of $\mathfrak{sl}_2$ that S is transverse to $\mathbb{O}$. In the sequel it will be convenient for us to consider S as an affine subspace in $\mathfrak{g}^*$ via the identification $\mathfrak{g}\cong\mathfrak{g}^*$. In particular, $\chi\in S$.

We need an action of the one-dimensional torus $\mathbb{K}^\times$ on $\mathfrak{g}^*$ that stabilizes S and contracts it to χ. Namely, the $\mathfrak{sl}_2$-triple (e,h,f) defines a homomorphism $\mathrm{SL}_2(\mathbb{K})\to G$. The group $\mathbb{K}^\times$ is embedded into $\mathrm{SL}_2(\mathbb{K})$ via $t\mapsto\operatorname{diag}(t,t^{-1})$. Composing these two homomorphisms we get a homomorphism (in fact, an embedding) $\gamma:\mathbb{K}^\times\to G$. For $\xi\in\mathfrak{g}(i)$ we have $\gamma(t).\xi=t^i\xi$. Consider a $\mathbb{K}^\times$-action (called the Kazhdan action) on $\mathfrak{g}^*$ given by $t\cdot\alpha=t^{-2}\gamma(t)\alpha$. This action fixes χ. Also it is easy to see that it preserves S. Finally, the representation theory of $\mathfrak{sl}_2$ implies that $\ker\operatorname{ad}(f)\subset\bigoplus_{i\leqslant 0}\mathfrak{g}(i)$. It follows that the action of $\mathbb{K}^\times$ contracts S to χ: $\lim_{t\to\infty}t.s=\chi$ for any $s\in S$. The contraction property has several very nice corollaries. For example, S intersects an adjoint orbit $\mathbb{O}'$ if and only if $\mathbb{O}\subset\overline{\mathbb{K}^\times\mathbb{O}'}$ and in this case the intersection $S\cap\mathbb{O}'$ is transversal.

The Kazhdan action gives rise to a (positive) grading on the algebra $\mathbb{K}[S]$ of regular functions on S. The following result was essentially obtained by Premet, [50], Theorem 4.6 (Kostant and Lynch also proved this in the special cases they considered).

Theorem 2.1. *There is an isomorphism* $\operatorname{gr}U(\mathfrak{g},e)\cong\mathbb{K}[S]$ *of graded algebras.*

As was shown by Gan and Ginzburg, [25], this result is a manifestation of the "quantization commutes with reduction" principle, see Subsection 2.5 for details.

2.5. Ramifications.

First, let us mention the work of Gan and Ginzburg, [25], where they gave a ramification of Premet's definition showing, in particular, that $U(\mathfrak{g},e)$ does not depend on the choice of $l\subset\mathfrak{g}(-1)$.

Namely, let $l\subset\mathfrak{g}(-1)$ be an arbitrary isotropic subspace of $\mathfrak{g}(-1)$ (e.g., $\{0\}$). Let $l^\angle$ denote the skew-orthogonal complement to l in $\mathfrak{g}(-1)$. Set $\mathfrak{m}^l:=l\oplus\bigoplus_{i\leqslant -2}\mathfrak{g}(i),\mathfrak{n}^l:=l^\angle\oplus\bigoplus_{i\leqslant -2}\mathfrak{g}(i)$. Then $\mathfrak{m}^l\subset\mathfrak{n}^l$, $\mathfrak{n}^l$ consists of nilpotent

elements, and $\langle\chi,[\mathfrak{m}_l,\mathfrak{n}_l]\rangle=0$. Let N^l be the connected subgroup of G with Lie algebra $\mathfrak{n}^l$. Then $\mathfrak{m}^l$ and the character $\chi:\mathfrak{m}^l\to\mathbb{K}$ are stable under the adjoint action of N^l. So N^l acts naturally on the quotient $\mathcal{U}/\mathcal{U}\mathfrak{m}^l_\chi$, where $\mathfrak{m}^l_\chi:=\{\xi-\langle\chi,\xi\rangle,\xi\in\mathfrak{m}^l\}$. Let $U(\mathfrak{g},e)^l:=(\mathcal{U}/\mathcal{U}\mathfrak{m}^l_\chi)^{N^l}$ be the space of invariants. It is easy to check that it has a natural algebra structure. It also has a Kazhdan filtration $\mathrm{K}_i\,U(\mathfrak{g},e)^l$, compare with the previous subsection.

Now let us remark that for $l_1\subset l_2$ we have a natural $\mathcal{U}$-module homomorphism $\mathcal{U}/\mathcal{U}\mathfrak{m}^{l_1}_\chi\to\mathcal{U}/\mathcal{U}\mathfrak{m}^{l_2}_\chi$ that gives rise to a filtered algebra homomorphism $U(\mathfrak{g},e)^{l_1}\to U(\mathfrak{g},e)^{l_2}$. It turns out that the latter is an isomorphism.

Also Gan and Ginzburg gave a very transparent explanation of an isomorphism $\operatorname{gr}U(\mathfrak{g},e)\cong\mathbb{K}[S]$. Namely, consider the restriction map $\pi:\mathfrak{g}^*\to\mathfrak{m}^{l*}$. The affine subspace $\pi^{-1}(\chi|_{\mathfrak{m}^l})\subset\mathfrak{g}^*$ is N^l-stable. Also it is easy to see that $S\subset\pi^{-1}(\chi|_{\mathfrak{m}^l})$. So we can consider a morphism $N^l\times S\to\pi^{-1}(\chi|_{\mathfrak{m}^l}),(n,s)\mapsto ns$. According to [25], this is an isomorphism (of algebraic varieties). Therefore $\mathbb{K}[S]$ gets identified with the classical Hamiltonian reduction $(S(\mathfrak{g})/S(\mathfrak{g})\mathfrak{m}^l_\chi)^{N^l}$.

Another ramification of the original definition of $U(\mathfrak{g},e)$ comes from the notion of a good grading on $\mathfrak{g}$, [20]. A grading $\mathfrak{g}=\bigoplus_{i\in\mathbb{Z}}\mathfrak{g}(i)$ is said to be *good* for e if $e\in\mathfrak{g}(2)$ and $\ker\operatorname{ad}(e)\subset\bigoplus_{i\geqslant 0}\mathfrak{g}(i)$. For instance, the grading given by (2.1) is good. For a comprehensive study of good gradings see [20].

Given a good grading on $\mathfrak{g}$, one constructs $\mathfrak{m}\subset\mathfrak{g}$ and defines $U(\mathfrak{g},e)$ using $\mathfrak{m}$ analogously to the above. The algebra $U(\mathfrak{g},e)$ does not depend on the choice of a good grading. This was first proved in [9].

The definition involving an arbitrary good grading is often useful. For example, one can sometimes find an *even* good grading when e is not even itself and embed $U(\mathfrak{g},e)$ into $U(\mathfrak{p})$ for an appropriate parabolic subalgebra $\mathfrak{p}\subset\mathfrak{g}$, compare with Subsection 2.2. This is always the case when $\mathfrak{g}\cong\mathfrak{sl}_n$, see [12], Introduction.

Also it is worth mentioning that there is a related definition of $U(\mathfrak{g},e)$ via the BRST quantization procedure which was used by physicists in the 90-s, see [6]. The proof that the BRST definition is equivalent to the one given above was obtained in [16]. See also [60], Section 3.

2.6. Additional properties of $U(\mathfrak{g},e)$. We want to make a few remarks about other properties of $U(\mathfrak{g},e)$.

Recall that $\mathcal{Z}$ stands for the center of $\mathcal{U}$. Restricting the natural map $\mathcal{U}^{\operatorname{ad}\mathfrak{m}}\to U(\mathfrak{g},e)$ to $\mathcal{Z}\subset\mathcal{U}^{\operatorname{ad}\mathfrak{m}}$, we get an algebra homomorphism $\mathcal{Z}\to U(\mathfrak{g},e)$. By [50], 6.2, this homomorphism is an embedding. It is clear that the image of $\mathcal{Z}$ lies in the center of $U(\mathfrak{g},e)$. Further, according to the footnote to Question 5.1 in [51], the image of $\mathcal{Z}$ actually coincides with the center of $U(\mathfrak{g},e)$ (Premet attributes the proof to Ginzburg). This generalizes Kostant's result mentioned in Subsection 2.1.

Also we remark that there is a natural action of the group $Q:=Z_G(e,h,f)$ on $U(\mathfrak{g},e)$. Namely, take $l=\{0\}$ in the Gan and Ginzburg definition. Then Q stabilizes both $\mathfrak{m}^l_\chi$ and N^l and so acts on $U(\mathfrak{g},e)^l$. Let $\mathfrak{q}$ stand for the Lie

algebra of Q. In [51] Premet constructed a Lie algebra embedding $\mathfrak{q} \hookrightarrow U(\mathfrak{g}, e)$ such that the adjoint action of $\mathfrak{q}$ on $U(\mathfrak{g}, e)$ coincides with the differential of the Q-action.

3. W-algebras Via Deformation Quantization

In this section we review the definition of W-algebras from [37]. It is based on Deformation quantization: a W-algebra is realized as an algebra of G-invariants in a quantization of a certain affine symplectic G-variety (called an *equivariant Slodowy slice*). In Subsection 3.1 we briefly explain generalities on star-products and on Fedosov's method to construct them. In Subsection 3.2 we present constructions of equivariant Slodowy slices and of W-algebras. Finally, in Subsection 3.3 we present a very important basic result on W-algebras, the *decomposition theorem*.

3.1. Fedosov quantization. In this subsection X is a smooth affine variety equipped with a symplectic form ω. Let $\{\cdot,\cdot\}$ denote the Poisson bracket on $\mathbb{K}[X]$ induced by ω. Let a reductive group $\widetilde{G}$ act on X preserving ω. By ξ_X we denote the image of $\xi \in \widetilde{\mathfrak{g}}$ under the homomorphism $\widetilde{\mathfrak{g}} \to \mathrm{Der}(\mathbb{K}[X])$ induced by the action.

We suppose that the $\widetilde{G}$-action is Hamiltonian, that is, admits a *moment map* $\mu : X \to \widetilde{\mathfrak{g}}^*$, i.e., a $\widetilde{G}$-equivariant morphism having the following property: for $H_\xi := \mu^*(\xi), \xi \in \widetilde{\mathfrak{g}}$, we have $\{H_\xi, \cdot\} = \xi_X$. Finally, we suppose that X is equipped with a $\mathbb{K}^\times$-action that commutes with $\widetilde{G}$ and satisfies $t.\omega = t^2\omega, t.H_\xi = t^2 H_\xi$ for all $t \in \mathbb{K}^\times, \xi \in \widetilde{\mathfrak{g}}$. We will present examples of this situation below.

By a *star-product* on $\mathbb{K}[X]$ (or on X) we mean a $\mathbb{K}$-bilinear map $* : \mathbb{K}[X] \times \mathbb{K}[X] \to \mathbb{K}[X][[\hbar]], (f,g) \mapsto f * g := \sum_{i=0}^{\infty} D_i(f,g)\hbar^{2i}$ satisfying the following axioms:

(a) The associativity axiom: a natural extension of $*$ to a $\mathbb{K}[[\hbar]]$-bilinear map $\mathbb{K}[X][[\hbar]] \times \mathbb{K}[X][[\hbar]] \to \mathbb{K}[X][[\hbar]]$ is an associative product, and $1 \in \mathbb{K}[X] \subset \mathbb{K}[X][[\hbar]]$ is a unit for $*$.

(b) The compatibility axiom: $D_0(f,g) = fg, D_1(f,g) - D_1(g,f) = \{f,g\}$. Equivalently, $f * g \equiv fg \mod \hbar^2$ and $[f,g] \equiv \hbar^2\{f,g\} \mod \hbar^4$.

(c) The locality axiom: D_i is a bidifferential operator of order at most i (i.e., for any fixed f the map $\mathbb{K}[X] \to \mathbb{K}[X] : g \mapsto D_i(f,g)$, is a differential operator of order at most i, and the same for any fixed g).

When we consider $\mathbb{K}[X][[\hbar]]$ as an algebra with respect to the star-product, we call it a *quantum algebra*.

We remark that the usual definition of a star-product looks like $f * g = \sum_{i=0}^{\infty} D_i(f,g)\hbar^i$ and in our definition we have $\hbar^2$ instead of $\hbar$. The reason for

this ramification is that our version is better compatible with the *Rees algebra construction*. This construction allows to pass from filtered $\mathbb{K}$-algebras to graded $\mathbb{K}[\hbar]$-algebras.

We will also need $*$ to be compatible with the $\widetilde{G}$- and $\mathbb{K}^\times$- actions on X.

(d) $\widetilde{G}$-invariance: $D_i : \mathbb{K}[X] \otimes \mathbb{K}[X] \to \mathbb{K}[X]$ is $\widetilde{G}$-equivariant.

(e) Homogeneity: D_i has degree $-2i$ with respect to $\mathbb{K}^\times$: i.e., for $f, g \in \mathbb{K}[X]$ of degrees j, k the element $D_i(f, g)$ has degree $k + j - 2i$.

Under the conditions (d) and (e), the product $\widetilde{G} \times \mathbb{K}^\times$ acts on $\mathbb{K}[X][[\hbar]]$ by automorphisms with $g.\hbar = \hbar, t.\hbar = t\hbar$ for all $g \in G, t \in \mathbb{K}^\times$.

It turns out that a star-product on X satisfying additionally (d) and (e) always exists. It is provided, for example, by Fedosov's construction, [21],[22]. Fedosov constructed a star-product on a C^∞-manifold starting from a *symplectic connection* ∇ and a closed $\mathbb{K}[[\hbar^2]]$-valued form Ω. By definition, a symplectic connection is a torsion-free connection on the tangent bundle such that the symplectic form is flat. Fedosov's construction can be carried over to the algebraic setting as long as a variety in consideration admits a symplectic connection. Since X is affine, this is the case, and, moreover, one can, in addition, assume that a symplectic connection is $\widetilde{G} \times \mathbb{K}^\times$-invariant, see [37], Proposition 2.2.2. For our purposes, it will be enough to consider the original construction from [21], where Ω is not used (i.e., equals 0).

The following proposition follows from results of Fedosov, see [38], Theorem 2.1.2 for details.

Proposition 3.1. *Let X be as above, and ∇ be a $\widetilde{G} \times \mathbb{K}^\times$-invariant symplectic connection on X. Further, let $*$ be the star-product produced from ∇ by the Fedosov construction. Then $*$ is $\widetilde{G}$-invariant and homogeneous. Moreover, the map $\xi \mapsto H_\xi$ is a* quantum comoment map *for the $\widetilde{G}$-action on $\mathbb{K}[X][[\hbar]]$, i.e., $\frac{1}{\hbar^2}[H_\xi, f] = \xi_X f$ for all $f \in \mathbb{K}[X][[\hbar]], \xi \in \widetilde{\mathfrak{g}}$.*

Also, according to Fedosov, $*$ does not depend on the choice of ∇ up to a suitably understood isomorphism, see, for example, [37], Proposition 2.2.5, for details.

Let us consider two standard examples.

The first example is easy. Let V be a vector space equipped with a non-degenerate form $\omega \in \bigwedge^2 V^*$. Let $\widetilde{G}$ act on V via a homomorphism $\widetilde{G} \to \mathrm{Sp}(V)$. Pick a homomorphism $\beta : \mathbb{K}^\times \to \mathrm{Sp}(V)^{\widetilde{G}}$ and define a $\mathbb{K}^\times$-action on V^* by $t.\alpha = t^{-1}\beta(t)\alpha$. So we get a symplectic variety $X = V^*$ equipped with a $\widetilde{G} \times \mathbb{K}^\times$-action satisfying the assumptions above with the moment map given by $\langle \mu(v), \xi \rangle = \frac{1}{2}\omega(\xi v, v)$. The algebra $\mathbb{K}[V^*]$ has a standard star-product called the *Moyal-Weyl* product. Namely, for $f, g \in \mathbb{K}[V^*]$ set $f * g := m(\exp(\frac{\omega}{2}\hbar^2)f \otimes g)$. Here $m : \mathbb{K}[V^*] \otimes \mathbb{K}[V^*] \to \mathbb{K}[V^*]$ stands for the multiplication map, while $\omega \in \bigwedge^2 V^*$ is assumed to act on $\mathbb{K}[V^*] \otimes \mathbb{K}[V^*]$ via contraction. The quantum

algebra $\mathbb{K}[V^*][\hbar]$ is naturally identified with the "homogeneous" version $\mathbf{A}_\hbar$ of the Weyl algebra of V, $\mathbf{A}_\hbar := T(V)[\hbar]/(u\otimes v - v\otimes u - \hbar^2\omega(u,v), u,v\in V)$.

Our second example is more involved although is also standard.

Let G be a connected reductive algebraic group. The cotangent bundle $X := T^*G$ of G is equipped with a canonical symplectic form ω. Set $\widetilde{G} := G\times G$ and consider the $\widetilde{G}$-action on X induced from the two-sided action of $\widetilde{G}$ on G. In more detail, we can identify T^*G with $G\times\mathfrak{g}^*$ using the trivialization by left-invariant forms. Then the "left" action of G on X is given by $g.(g_1,\alpha) = (gg_1,\alpha)$, while the "right" action is $g.(g_1,\alpha) = (g_1g^{-1}, g.\alpha)$. Finally, let $\mathbb{K}^\times$ act on X by $t.(g_1,\alpha) = (g_1, t^{-2}\alpha)$. Clearly, ω is $\widetilde{G}$-invariant and $t.\omega = t^2\omega$. A moment map $\mu: X\to\widetilde{\mathfrak{g}}^* = \mathfrak{g}^*\oplus\mathfrak{g}^*$ is given by $(g,\alpha)\mapsto(g.\alpha,\alpha)$.

Pick a $\widetilde{G}\times\mathbb{K}^\times$-invariant connection ∇ on X and produce the star-product $*$ from ∇. From the grading considerations, we see that $\mathbb{K}[X][\hbar]$ is a subalgebra in the quantum algebra $\mathbb{K}[X][[\hbar]]$.

There is a standard alternative description of $\mathbb{K}[X][\hbar]$, see, for example, Subsection 7.1 of [40]. Consider the algebra $\mathcal{D}(G)$ of linear differential operators on G. Let $\mathrm{F}_i\,\mathcal{D}(G)$ be the space of differential operators of order $\leqslant i/2$. Consider the Rees algebra $\mathcal{D}_\hbar(G) := \bigoplus_{i=0}^\infty \mathrm{F}_i\,\mathcal{D}(G)\hbar^i \subset \mathcal{D}(G)[\hbar]$ of $\mathcal{D}(G)$. Then there is a $\widetilde{G}\times\mathbb{K}^\times$-equivariant isomorphism $\mathbb{K}[X][\hbar]\cong\mathcal{D}_\hbar(G)$ of $\mathbb{K}[\hbar]$-algebras.

Taking the G-invariants in the algebra $\mathbb{K}[T^*G][\hbar]\cong\mathcal{D}_\hbar(G)$ (say for the left G-action), we get a new (star-)product on $\mathbb{K}[\mathfrak{g}^*][\hbar] = \mathbb{K}[T^*G][\hbar]^G$. But $\mathcal{D}_\hbar(G)^G$ is nothing else but a homogeneous version $\mathcal{U}_\hbar$ of the universal enveloping algebra $\mathcal{U}$ of $\mathfrak{g}$, $\mathcal{U}_\hbar := T(\mathfrak{g})[\hbar]/(\xi\otimes\eta-\eta\otimes\xi-\hbar^2[\xi,\eta], \xi,\eta\in\mathfrak{g})$. In the next subsection we will use a similar recipe to define a W-algebra.

3.2. Equivariant Slodowy slices and W-algebras.

A variety we need in the approach to W-algebras from [37] is as follows. Recall the Slodowy slice $S\subset\mathfrak{g}^*$, Subsection 2.4. Set $X := G\times S\subset G\times\mathfrak{g}^* = T^*G$. The variety X is called the *equivariant Slodowy slice*. Clearly, $X\subset T^*G$ is stable with respect to the left G-action. Also it is stable under the restriction of the right G-action to $Q = Z_G(e,h,f)$. Finally, X is stable under a *Kazhdan* $\mathbb{K}^\times$-action given by $t.(g,\alpha) = (g\gamma(t)^{-1}, t^{-2}\gamma(t)\alpha)$, where $\gamma:\mathbb{K}^\times\to G$ was introduced in Subsection 2.4. Consider the 2-form ω on X obtained by the restriction of the natural symplectic form from T^*G. One can show that ω is non-degenerate. So X becomes a symplectic variety. It satisfies the assumptions in the beginning of the previous subsection with $\widetilde{G} := G\times Q$, the Kazhdan action of $\mathbb{K}^\times$ and a moment map $X\to\mathfrak{g}^*\oplus\mathfrak{q}^*$ restricted from T^*G.

Pick a $\widetilde{G}\times\mathbb{K}^\times$-invariant symplectic connection ∇ on X and produce a star-product $f*g = \sum_{i=0}^\infty D_i(f,g)\hbar^{2i}$, using the Fedosov construction. [37], Proposition 2.1.5 implies that $\mathbb{K}[X][\hbar]\subset\mathbb{K}[X][[\hbar]]$ is closed with respect to the star-product. We call the quantum algebra $\mathbb{K}[X][\hbar]$ a *homogeneous equivariant W-algebra* and denote it by $\widetilde{\mathcal{W}}_\hbar$. A *homogeneous W-algebra* is, by definition, $\mathcal{W}_\hbar := \widetilde{\mathcal{W}}_\hbar^G$. Finally, define a W-algebra $\mathcal{W}$ as $\mathcal{W}_\hbar/(\hbar-1)$. So, as a vector space $\mathcal{W}$ is the same as $\mathbb{K}[S]$ but the product on $\mathcal{W}$ is given by $fg := \sum_{i=0}^\infty D_i(f,g)$.

The algebra $\mathcal{W}$ comes equipped with

- a filtration $\mathrm{F}_i\,\mathcal{W}$ induced from the grading on $\mathcal{W}_\hbar$.
- an action of Q.
- a homomorphism (in fact an embedding) $\mathfrak{q} \hookrightarrow \mathcal{W}$ of Lie algebras such that the adjoint action of $\mathfrak{q}$ on $\mathcal{W}$ coincides with the differential of the Q-action.
- a homomorphism $\mathcal{Z} \to \mathcal{W}$ (induced from the quantum comoment map $\mathfrak{g} \to \widetilde{\mathcal{W}_\hbar}$).

It turns out that $\mathcal{W}$ is isomorphic to $U(\mathfrak{g}, e)$. More precisely, we have the following result.

Theorem 3.2 ([37], Corollary 3.3.3). *There is a filtration preserving isomorphism $\mathcal{W} \to U(\mathfrak{g}, e)$.*

One can prove, in addition, that this isomorphism is Q-equivariant (although this is not written down explicitly) and intertwines the homomorphisms $\mathcal{Z} \to \mathcal{W}, U(\mathfrak{g}, e)$ (this is proved in [38], the end of Subsection 2.2).

3.3. Decomposition theorem. Let x denote the point $(1, \chi) \in X \subset T^*G = G \times \mathfrak{g}^*$. We remark that the orbit Gx is closed (as any orbit in T^*G) and also $Q \times \mathbb{K}^\times$-stable. Consider the formal neighborhoods $(T^*G)^\wedge_{Gx}, X^\wedge_{Gx}$ of Gx in T^*G and X and the formal neighborhood $(V^*)^\wedge_0$ of 0 in V^*. Being defined by bidifferential operators, the star-products on $\mathbb{K}[T^*G][\hbar], \mathbb{K}[X][\hbar], \mathbb{K}[V^*][\hbar]$ extend to the corresponding completions $\mathbb{K}[T^*G]^\wedge_{Gx}[[\hbar]], \mathbb{K}[X]^\wedge_{Gx}[[\hbar]], \mathbf{A}^\wedge_\hbar := \mathbb{K}[V^*]^\wedge_0[[\hbar]]$.

Taking the G-invariants in $\mathbb{K}[T^*G]^\wedge_{Gx}[[\hbar]], \mathbb{K}[X]^\wedge_{Gx}[[\hbar]]$ we get star-products on the completions $\mathcal{U}^\wedge_\hbar := \mathbb{K}[\mathfrak{g}^*]^\wedge_\chi, \mathcal{W}^\wedge_\hbar := \mathbb{K}[S]^\wedge_\chi[[\hbar]]$. The algebras $\mathcal{U}^\wedge_\hbar, \mathcal{W}^\wedge_\hbar, \mathbf{A}^\wedge_\hbar$ come equipped with natural (complete and separated) topologies. We remark that the completions $\mathcal{U}^\wedge_\hbar, \mathcal{W}^\wedge_\hbar, \mathbf{A}^\wedge_\hbar$ can be defined completely algebraically, as the inverse limits of $\mathcal{U}_\hbar, \mathcal{W}_\hbar, \mathbf{A}_\hbar$ with respect to the powers of appropriate maximal ideals, see [38], Subsection 2.4 for details.

The following theorem follows from [37], Theorem 3.3.1.

Theorem 3.3. *There is a $Q \times \mathbb{K}^\times$-equivariant isomorphism $\Phi_\hbar : \mathcal{U}^\wedge_\hbar \to \mathbf{A}^\wedge_\hbar \widehat{\otimes}_{\mathbb{K}[[\hbar]]} \mathcal{W}^\wedge_\hbar$ of topological $\mathbb{K}[[\hbar]]$-algebras.*

Here $\widehat{\otimes}$ stands for the completed tensor product: we take the usual tensor product of topological $\mathbb{K}[[\hbar]]$-algebras and then complete it with respect to the induced topology.

Theorem 3.3 is extremely important in the study of W-algebras. It can be used to prove Theorem 3.2, to prove the category equivalence theorems 4.1,4.3 in the next section, and also to relate the sets of two-sided ideals of $\mathcal{U}$ and of $\mathcal{W}$, see Section 5.

4. Category Equivalences

This section is devoted to the description of two category equivalences between suitable categories of $\mathcal{W}$-modules and of $\mathcal{U}$-modules. In the first subsection we recall an equivalence proved by Skryabin in [57]. This is an equivalence between the category of all $\mathcal{W}$-modules and the category of *Whittaker* $\mathcal{U}$-modules. Then we discuss some corollaries of Skryabin's theorem, in particular, a localization theorem due to Ginzburg, [26]. Subsection 4.2 deals with a ramification of Skryabin's equivalence conjectured in [10] and proved in [39]. This is an equivalence between the *category* $\mathcal{O}$ for a W-algebra and the category of *generalized Whittaker* $\mathcal{U}$-modules.

4.1. Whittaker modules and Skryabin's equivalence. Recall that in Subsection 2.3 we have defined the W-algebra $\mathcal{W} = U(\mathfrak{g}, e)$ as the quantum Hamiltonian reduction $(\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi)^{\mathrm{ad}\,\mathfrak{m}}$. In other words, $\mathcal{W} = \mathrm{End}_{\mathcal{U}}(\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi)^{op}$. In particular, $\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi$ is a $\mathcal{U}$-$\mathcal{W}$-bimodule.

We say that a $\mathcal{U}$-module M is *Whittaker* if the action of $\mathfrak{m}_\chi$ on M is locally nilpotent. For instance, $\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi$ is easily seen to be Whittaker. Whittaker modules form a Serre subcategory in the category $\mathcal{U}$-Mod. Denote the category of Whittaker $\mathcal{U}$-modules by Wh.

The bimodule $\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi$ gives rise to the following functors:

$$\begin{aligned}&\mathrm{Wh} \to \mathcal{W}\text{-}\,\mathrm{Mod} : M \mapsto \mathrm{Hom}_{\mathcal{U}}(\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi, M) = M^{\mathfrak{m}_\chi} := \\ &\qquad\qquad := \{m \in M : \xi m = \langle \chi, \xi\rangle m, \forall \xi \in \mathfrak{m}\}.\\ &\mathcal{W}\text{-}\,\mathrm{Mod} \to \mathrm{Wh} : N \mapsto \mathcal{U}/\mathcal{U}\mathfrak{m}_\chi \otimes_{\mathcal{W}} N.\end{aligned}$$

We denote the second functor by Sk.

The following important theorem was proved in [57].

Theorem 4.1. *The functors above are quasi-inverse equivalences.*

Let us mention several important corollaries of this theorem.

The Beilinson-Bernstein localization theorem, [4], is a crucial result in the representation theory of $\mathcal{U}$. There is an analog of this theorem for W-algebras due to Ginzburg, [26]. See also [18] for an alternative approach.

Recall the Beilinson-Bernstein theorem. Pick a Cartan subalgebra $\mathfrak{h} \subset \mathfrak{g}$. Let $\Delta \subset \mathfrak{h}^*$ be the root system, W the Weyl group, and $\Pi \subset \Delta$ be a system of simple roots. Recall the dot action of W on $\mathfrak{h}^*$ given by $w \cdot \lambda = w(\lambda + \rho) - \rho$, where, as usual, ρ stands for the half of the sum of all positive roots. The center $\mathcal{Z}$ of $\mathcal{U}$ gets identified via the Harish-Chandra isomorphism with the invariant algebra $\mathbb{K}[\mathfrak{h}^*]^W$, the invariants are taken with respect to the dot action.

To any $\lambda \in \mathfrak{h}^*$ one assigns a sheaf $\mathcal{D}_\lambda$ of twisted differential operators on the flag variety $\mathcal{B}$ of G, see [4]. The algebra $\Gamma(\mathcal{B}, \mathcal{D}_\lambda)$ of global sections is naturally identified with the quotient $\mathcal{U}_\lambda := \mathcal{U}/\mathcal{U}I_\lambda$, where I_λ denotes the maximal ideal of $W \cdot \lambda$ in $\mathcal{Z}$. So to a $\mathcal{D}_\lambda$-module M one can assign the $\mathcal{U}_\lambda$-module $\Gamma(\mathcal{B}, M)$. The

functor $\Gamma(\mathcal{B},\bullet)$ has a left adjoint: the localization functor $\mathcal{D}_\lambda\otimes_{\Gamma(\mathcal{B},\mathcal{D}_\lambda)}\bullet$. The Beilinson-Bernstein theorem states that the functor $\Gamma(\mathcal{B},\bullet)$ is an equivalence provided λ is regular and dominant, i.e., $\langle\lambda+\rho,\alpha\rangle\notin\mathbb{Z}_{\leqslant 0}$ for any $\alpha\in\Delta$.

Let us explain some details on Ginzburg's localization theorem. For more details the reader is referred to [26].

One can consider the sheaf $\mathcal{D}_\lambda$ as a quantization of the symplectic variety $T^*\mathcal{B}$. An analog of $T^*\mathcal{B}$ for $\mathcal{W}$ is the *Slodowy variety* defined as follows. The action of G on $T^*\mathcal{B}$ is Hamiltonian, the Springer resolution morphism $\mu: T^*\mathcal{B}\to\mathfrak{g}^*$ is a moment map. Recall the projection $\pi:\mathfrak{g}^*\twoheadrightarrow\mathfrak{m}^*$. Then $\pi\circ\mu$ is a moment map for the M-action on $T^*\mathcal{B}$. By definition, the Slodowy variety $\mathcal{S}$ is the Hamiltonian reduction $(\pi\circ\mu)^{-1}(\chi|_{\mathfrak{m}})/M$.

To define an analog of the sheaf $\mathcal{D}_\lambda$ in the W-algebra setting Ginzburg uses the language of *directed algebras* (one can also use the language of microlocal sheaves, see [18]). Once this analog is defined the Beilinson-Bernstein theorem transfers to the W-algebra setting verbatim. The scheme of the proof is as follows: one introduces the notion of a Whittaker $\mathcal{D}_\lambda$-module, shows that the functors in the Beilinson-Bernstein theorem restrict to equivalences between the Whittaker subcategories, and then uses the Skryabin theorem.

A related development is as follows. Let L be a finite dimensional $\mathfrak{g}$-module, and M be a Whittaker $\mathfrak{g}$-module. Then $L\otimes M$ is again a Whittaker $\mathfrak{g}$-module. This allows to define tensor products of finite dimensional $\mathfrak{g}$-modules with $\mathcal{W}$-modules. These tensor products are studied in detail in [27].

4.2. Category $\mathcal{O}$ for W-algebras. In the representation theory of $\mathcal{U}$ a crucial role is played by the category $\mathcal{O}$ established by Bernstein, I. Gelfand and S. Gelfand in [5]. There is an analog of the BGG category $\mathcal{O}$ for $\mathcal{W}$ introduced by Brundan, Goodwin and Kleshchev in [10]. The most important result about this category is that it is equivalent to a certain category of *generalized Whittaker* $\mathcal{U}$-modules, [39]. Our exposition follows [39].

Recall the group $Q:=Z_G(e,h,f)$ acting on $\mathcal{W}$ and an embedding $\mathfrak{q}\hookrightarrow\mathcal{W}$, see Subsections 2.6,3.2. Pick a Cartan subalgebra $\mathfrak{t}\subset\mathfrak{q}$ and set $\mathfrak{l}:=\mathfrak{z}_{\mathfrak{g}}(\mathfrak{t})$. Then $\mathfrak{l}$ is a minimal Levi subalgebra in $\mathfrak{g}$ containing e. Further, pick an integral (=lying in the character lattice of the corresponding maximal torus of Q) element $\theta\in\mathfrak{t}$ with $\mathfrak{z}_{\mathfrak{g}}(\theta)=\mathfrak{l}$. A category we are going to consider will depend on θ.

Consider the decomposition $\mathcal{W}=\bigoplus_{i\in\mathbb{Z}}\mathcal{W}_i$, where $\mathcal{W}_i:=\{w\in\mathcal{W}|[\theta,w]=iw\}$. Set $\mathcal{W}_{\geqslant 0}:=\bigoplus_{i\geqslant 0}\mathcal{W}_i,\mathcal{W}_{>0}:=\bigoplus_{i>0}\mathcal{W}_i,\mathcal{W}^+_{\geqslant 0}:=\mathcal{W}_{\geqslant 0}\cap\mathcal{W}\mathcal{W}_{>0}$. Then $\mathcal{W}_{\geqslant 0}$ is a subalgebra in $\mathcal{W}$, while $\mathcal{W}_{>0}$ and $\mathcal{W}^+_{\geqslant 0}$ are two-sided ideals in $\mathcal{W}_{\geqslant 0}$.

We say that a $\mathcal{W}$-module N belongs to the category $\mathcal{O}(\theta)$ (in [39] this category was denoted by $\widetilde{\mathcal{O}}^{\mathfrak{l}}(\theta)$) if

- N is finitely generated.
- $\mathfrak{t}\subset\mathcal{W}$ acts on N by diagonalizable endomorphisms.
- $\mathcal{W}_{>0}$ acts on N by locally nilpotent endomorphisms.

For example, when e is distinguished (i.e., $\mathfrak{q} = \{0\}$), then $\mathcal{O}(\theta)$ consists precisely of all finite dimensional $\mathcal{W}$-modules. In this case the notion of the category $\mathcal{O}$ is pretty useless. The other extreme is the case when e is principal in $\mathfrak{l}$. We will see below that in this case we can say a lot about $\mathcal{O}(\theta)$.

Let us present an important construction of a module in $\mathcal{O}(\theta)$. Pick a $\mathcal{W}_{\geqslant 0}/\mathcal{W}^+_{\geqslant 0}$-module N^0 with diagonalizable $\mathfrak{t}$-action (e.g., irreducible). Define the *Verma* module $\Delta^\theta(N^0)$ by $\Delta^\theta(N^0) := \mathcal{W} \otimes_{\mathcal{W}_{\geqslant 0}} N^0$.

The properties of $\mathcal{O}(\theta)$ are quite expectable.

Proposition 4.2. *1. If N^0 is irreducible, then $\Delta^\theta(N^0)$ has a unique irreducible quotient, say $L^\theta(N^0)$.*

2. Any irreducible module in $\mathcal{O}(\theta)$ is isomorphic to $L^\theta(N^0)$ for unique N^0.

3. Let $N \in \mathcal{O}(\theta)$ be such that all $\mathfrak{t}$-eigenspaces in N are finite dimensional. Then N has finite length.

4. $\Delta^\theta(N^0)$ with $\dim N^0 < \infty$ satisfies the assumptions of (3).

This is proved in [10], Theorem 4.5, Corollary 4.12 (in [10] a bit different definition was used, in particular, the assumption in (3) was a part of the definition, but this does not matter).

The most crucial property of $\mathcal{O}(\theta)$ is that it is equivalent to a certain category of $\mathcal{U}$-modules. To define this category we need some more notation.

Let $\mathfrak{g} = \bigoplus_{i\in\mathbb{Z}} \mathfrak{g}_i$ be the decomposition into the eigenspaces of $\operatorname{ad}\theta$. In particular, $\mathfrak{l} = \mathfrak{g}_0$. Form the subalgebra $\underline{\mathfrak{m}} \subset \mathfrak{g}_0$ by analogy with $\mathfrak{m} \subset \mathfrak{g}$ but using the pair $(\mathfrak{g}_0, e)$ instead of $(\mathfrak{g}, e)$. We define the W-algebra $\mathcal{W}^0 := U(\mathfrak{g}_0, e)$. This notation is different from [39] but agrees with [40]. Consider the subalgebra $\widetilde{\mathfrak{m}} := \underline{\mathfrak{m}} \oplus \mathfrak{g}_{>0} \subset \mathfrak{g}$ (where $\mathfrak{g}_{>0} := \bigoplus_{i>0} \mathfrak{g}_i$) and set $\widetilde{\mathfrak{m}}_\chi := \{\xi - \langle\chi, \xi\rangle, \xi \in \widetilde{\mathfrak{m}}\}$. The element $\chi \in \mathfrak{g}^*$ is $\mathfrak{t}$-invariant and so vanishes on $\mathfrak{g}_{>0}$. Hence $\widetilde{\mathfrak{m}}_\chi = \underline{\mathfrak{m}}_\chi \times \mathfrak{g}_{>0}$.

We say that a $\mathcal{U}$-module M is a *generalized Whittaker* module (for e and θ) if

- M is finitely generated.
- $\mathfrak{t}$ acts on M by diagonalizable endomorphisms.
- $\widetilde{\mathfrak{m}}_\chi$ acts on M by locally nilpotent endomorphisms.

The category of generalized Whittaker modules will be denoted by $\mathrm{Wh}(\theta)$ (this notation is again different from the one used in [39]).

Again, one can define a Verma module in $\mathrm{Wh}(\theta)$. Let N^0 be a $\mathcal{W}^0$-module with diagonalizable $\mathfrak{t}$-action. Let $\mathrm{Sk}_0 : \mathcal{W}^0\text{-Mod} \to U(\mathfrak{g}_0)\text{-Mod}$ be the Skryabin functor (for the pair $\mathfrak{g}_0, e$). Define the Verma module $\Delta^{e,\theta}(N^0) := \mathcal{U} \otimes_{U(\mathfrak{g}_{\geqslant 0})} \mathrm{Sk}_0(N^0)$, where $U(\mathfrak{g}_{\geqslant 0})$ acts on $\mathrm{Sk}_0(N^0)$ via a natural epimorphism $U(\mathfrak{g}_{\geqslant 0}) \twoheadrightarrow U(\mathfrak{g}_0)$.

The following theorem is (a part of) the main result of [39].

Theorem 4.3 ([39], Theorem 4.1). *There is an isomorphism* $\Psi : \mathcal{W}^0 \to \mathcal{W}_{\geqslant 0}/\mathcal{W}^+_{\geqslant 0}$ *and an equivalence* $\mathcal{K} : \mathrm{Wh}(\theta) \to \mathcal{O}(\theta)$ *of abelian categories such that the functors* $\mathcal{K}(\Delta^{e,\theta}(\bullet))$ *and* $\Delta^{\theta}(\Psi_*(\bullet))$ *from the category of* $\mathfrak{t}$*-diagonalizable* $\mathcal{W}^0$*-modules to* $\mathcal{O}(\theta)$ *are isomorphic. Here* Ψ_* *denotes the push-forward functor with respect to the isomorphism* Ψ*.*

Let us make a remark on an isomorphism Ψ. Such an isomorphism was first established in [10]. It is however not completely clear if one can use the isomorphism from [10] in Theorem 4.3. A peculiar feature of both isomorphisms is that they do not intertwine the embeddings $\mathfrak{t} \hookrightarrow \mathcal{W}_{\geqslant 0}/\mathcal{W}^+_{\geqslant 0}, \mathcal{W}^0$ but rather induce a shift on $\mathfrak{t}$. Since this shift will be of importance later we will give some details, see Remark 5.5 in [39]. Namely, let ι^0, ι denote the embeddings of $\mathfrak{t}$ to $\mathcal{W}^0, \mathcal{W}_{\geqslant 0}/\mathcal{W}^+_{\geqslant 0}$, respectively. Then we have $\iota(\xi) = \Psi(\iota^0(\xi)) - \langle \delta, \xi \rangle$ for an element $\delta \in \mathfrak{t}^*$ constructed as follows. Pick a Cartan subalgebra $\mathfrak{h} \subset \mathfrak{g}$ containing $\mathfrak{t}$ and h. Let $\Delta_{<0}$ denote the set of all roots α of $\mathfrak{g}$ with $\langle \alpha, \theta \rangle < 0$. Set

$$\delta := \sum_{\alpha \in \Delta_{<0}, \langle \alpha, h \rangle = 1} \frac{1}{2} \alpha|_{\mathfrak{t}} + \sum_{\alpha \in \Delta_{<0}, \langle \alpha, h \rangle \geqslant 2} \alpha|_{\mathfrak{t}}. \tag{4.1}$$

Till the end of the subsection we consider the category $\mathrm{Wh}(\theta)$ in the special case when e is principal in $\mathfrak{l}$. Here $\mathrm{Wh}(\theta)$ (with a slightly different definition) was studied before by McDowell, [43], by Milicic and Soergel, [45], and by Backelin, [1].

To proceed we need some more notation. Choose a system Π of simple roots such that θ is dominant. Then $\Pi_0 := \{\alpha \in \Pi : \langle \alpha, \theta \rangle = 0\}$ is a system of simple roots for $\mathfrak{l}$. Let $\Delta_+, \Delta_{\mathfrak{l}+}$ denote the systems of positive roots for $\mathfrak{g}$ and $\mathfrak{l}$. For a root α let e_α denote a corresponding weight vector in $\mathfrak{g}$. Further, let $W_{\mathfrak{l}}$ denote the Weyl group of $\mathfrak{l}$. We have the dot action of W on $\mathfrak{h}^*$ defined as in the previous subsection.

The W-algebra $\mathcal{W}^0$ is identified with the center $Z(\mathfrak{l})$ of $U(\mathfrak{l})$. So all irreducible $\mathcal{W}^0$-module are 1-dimensional. The set of their isomorphism classes is in one-to-one correspondence with orbits of the dot action of $W_{\mathfrak{l}}$ on $\mathfrak{h}^*$.

One may assume that $e = \sum_{\alpha \in \Pi_0} e_{-\alpha}$. Then $\widetilde{\mathfrak{m}}$ is nothing else but the maximal nilpotent subalgebra $\mathfrak{n}$ of $\mathfrak{g}$ corresponding to Π. Also we have $\langle \chi, e_\alpha \rangle \neq 0$ if and only if $\alpha \in \Pi_0$. So we recover the setting of [1],[43],[45].

For $\lambda \in \mathfrak{h}^*$ let us write $\Delta(\lambda), L(\lambda)$ for the Verma and irreducible modules with highest weight λ in the BGG category $\mathcal{O}$ and $\Delta^{e,\theta}(\lambda), L^{e,\theta}(\lambda)$ for the Verma and irreducible modules in $\mathrm{Wh}(\theta)$ corresponding to $W_{\mathfrak{l}} \cdot \lambda$.

In [45] Milicic and Soergel proved that the (infinitesimal) block of $\mathrm{Wh}(\theta)$ corresponding to a regular integral central character is equivalent to the block of the BGG category $\mathcal{O}$ with certain *singular* integral character that can be recovered from Π_0. For a generalization of this equivalence to the parabolic setting see [61].

For other blocks in $\mathrm{Wh}(\theta)$ (corresponding to singular/non-integral central characters) the situation is more subtle. But still one can relate the multiplicities in $\mathcal{O}$ and in $\mathrm{Wh}(\theta)$. For $\lambda,\mu\in\mathfrak{h}^*$ let $[\Delta(\lambda):L(\mu)],[\Delta^{e,\theta}(\lambda):L^{e,\theta}(\mu)]$ denote the multiplicities in the corresponding categories.

Theorem 4.4 ([1], Theorem 6.2). *Let $\lambda,\mu\in\mathfrak{h}^*$. If*

1. *$\lambda\in W\cdot\mu$,*
2. *and there is $w\in W_{\mathfrak{l}}$ such that $w\cdot\mu$ is antidominant for $\mathfrak{l}$ and $\lambda-w\cdot\mu\in\mathrm{Span}_{\mathbb{Z}_{\geqslant 0}}(\Delta^+)$,*

then $[\Delta^{e,\theta}(\lambda):L^{e,\theta}(\mu)]=[\Delta(\lambda):L(w\cdot\mu)]$. Otherwise, $[\Delta^{e,\theta}(\lambda):L^{e,\theta}(\mu)]=0$.

An element $\lambda\in\mathfrak{h}^*$ is said to be antidominant for $\mathfrak{l}$ if $\langle\lambda+\rho,\alpha^\vee\rangle\notin\mathbb{Z}_{>0}$ for any $\alpha\in\Delta_{\mathfrak{l}+}$.

5. Ideals in $U(\mathfrak{g})$ Versus Ideals in $\mathcal{W}$

In this section we will construct maps between the sets $\mathfrak{Id}(\mathcal{U})$ and $\mathfrak{Id}(\mathcal{W})$ of two-sided ideals in $\mathcal{U}$ and $\mathcal{W}$, respectively. This is done in the first two subsections. In Subsection 5.3 we explain how these maps allow to relate (isomorphism classes of) finite dimensional irreducible $\mathcal{W}$-modules to primitive ideals $\mathcal{J}\subset\mathcal{U}$ such that the associated variety $\mathrm{V}(\mathcal{U}/\mathcal{J})$ is $\overline{\mathbb{O}}$. We conclude the section with some remarks in Subsection 5.4.

5.1. Map $\bullet_\dagger:\mathfrak{Id}(\mathcal{U})\to\mathfrak{Id}(\mathcal{W})$. Recall the algebras $\mathcal{U}_\hbar,\mathbf{A}_\hbar$ from Subsection 3.1, $\mathcal{W}_\hbar$ from Subsection 3.2, and the topological algebras $\mathcal{U}_\hbar^\wedge,\mathbf{A}_\hbar^\wedge,\mathcal{W}_\hbar^\wedge$ from Subsection 3.3. By Theorem 3.3, $\mathcal{U}_\hbar^\wedge\cong\mathbf{A}_\hbar^\wedge\widehat{\otimes}_{\mathbb{K}[[\hbar]]}\mathcal{W}_\hbar^\wedge$. Let us introduce suitable sets of ideals of $\mathcal{U}_\hbar,\mathcal{U}_\hbar^\wedge,\mathcal{W}_\hbar^\wedge,\mathcal{W}_\hbar$. Namely, let $\mathfrak{Id}_\hbar(\mathcal{U}_\hbar)$ denote the set of $\mathbb{K}^\times$-stable $\hbar$-saturated two-sided ideals in $\mathcal{U}_\hbar$ (an ideal $\mathcal{J}_\hbar\subset\mathcal{U}_\hbar$ is said to be $\hbar$-saturated if $\hbar a\in\mathcal{J}_\hbar$ implies $a\in\mathcal{J}_\hbar$, in other words, if the quotient $\mathcal{U}_\hbar/\mathcal{J}_\hbar$ is a flat $\mathbb{K}[\hbar]$-module). Define the sets $\mathfrak{Id}_\hbar(\mathcal{U}_\hbar^\wedge),\mathfrak{Id}_\hbar(\mathcal{W}_\hbar),\mathfrak{Id}_\hbar(\mathcal{W}_\hbar^\wedge)$ in a similar way. We define a map $\bullet_\dagger$ as the composition

$$\mathfrak{Id}(\mathcal{U})\xrightarrow{(a)}\mathfrak{Id}_\hbar(\mathcal{U}_\hbar)\xrightarrow{(b)}\mathfrak{Id}_\hbar(\mathcal{U}_\hbar^\wedge)\xrightarrow{(c)}\mathfrak{Id}_\hbar(\mathcal{W}_\hbar^\wedge)\xrightarrow{(d)}\mathfrak{Id}_\hbar(\mathcal{W}_\hbar)\xrightarrow{(e)}\mathfrak{Id}(\mathcal{W}). \tag{5.1}$$

Let us describe the intermediate maps.

(a): this map sends $\mathcal{J}\in\mathfrak{Id}(\mathcal{U})$ to $R_\hbar(\mathcal{J}):=\bigoplus(\mathcal{J}\cap\mathrm{F}_i\mathcal{U})\hbar^i$. It is a bijection, the inverse map sends $\mathcal{J}_\hbar\in\mathfrak{Id}_\hbar(\mathcal{U}_\hbar)$ to its image under the natural epimorphism $\mathcal{U}_\hbar\twoheadrightarrow\mathcal{U}_\hbar/(\hbar-1)=\mathcal{U}$.

(b): this map sends $\mathcal{J}_\hbar\in\mathfrak{Id}_\hbar(\mathcal{U}_\hbar)$ to its closure $\mathcal{J}_\hbar^\wedge\subset\mathcal{U}_\hbar^\wedge$. Equivalently, $\mathcal{J}_\hbar^\wedge=\mathcal{U}_\hbar^\wedge\mathcal{J}_\hbar$. This map is neither injective (but it is easy to say when two ideals have the same image) nor surjective.

(c): this map sends $\mathcal{J}'_\hbar\in\mathfrak{Id}_\hbar(\mathcal{U}_\hbar^\wedge)$ to $\mathcal{I}'_\hbar:=\mathcal{J}'_\hbar\cap\mathcal{W}_\hbar^\wedge$. It is a bijection: its inverse sends $\mathcal{I}'_\hbar\in\mathfrak{Id}_\hbar(\mathcal{W}_\hbar^\wedge)$ to $\mathbf{A}_\hbar^\wedge\widehat{\otimes}_{\mathbb{K}[[\hbar]]}\mathcal{I}'_\hbar$.

(d): this map sends $\mathcal{I}'_\hbar \in \mathfrak{Id}_\hbar(\mathcal{W}_\hbar^\wedge)$ to $\mathcal{I}_\hbar := \mathcal{I}'_\hbar \cap \mathcal{W}_\hbar$. It is again a bijection, its inverse sends $\mathcal{I}_\hbar \in \mathfrak{Id}_\hbar(\mathcal{W}_\hbar)$ to its closure.

(e): this map is analogous to the inverse of (a).

Proposition 5.1. *The map $\mathcal{J} \mapsto \mathcal{J}_\dagger$ has the following properties.*

(1) $\mathcal{J}_\dagger$ is Q-stable.

(2) $\operatorname{gr}\mathcal{W}/\mathcal{J}_\dagger$ is the pull-back of the $\mathbb{K}[\mathfrak{g}^*]$-module $\operatorname{gr}\mathcal{U}/\mathcal{J}$ to $S \subset \mathfrak{g}^*$.

(3) $\mathcal{J}_\dagger = \mathcal{W}$ if and only if $\mathbb{O} \cap \mathrm{V}(\mathcal{U}/\mathcal{J}) = \varnothing$.

(4) $\mathcal{J}_\dagger$ is a proper ideal of finite codimension in $\mathcal{W}$ if and only if $\overline{\mathbb{O}}$ is an irreducible component of $\mathrm{V}(\mathcal{U}/\mathcal{J})$. In this case $\dim \mathcal{W}/\mathcal{J}_\dagger$ equals the multiplicity of $\mathcal{U}/\mathcal{J}$ on $\mathbb{O}$.

(5) The natural map $(\mathcal{J}/\mathcal{J}\mathfrak{m}_\chi)^{\operatorname{ad}\mathfrak{m}} \to (\mathcal{U}/\mathcal{U}\mathfrak{m}_\chi)^{\operatorname{ad}\mathfrak{m}}$ is injective. Its image coincides with $\mathcal{J}_\dagger$.

(1) follows directly from the construction. (2) is Proposition 3.4.2 in [37]. (3) and (4) follow from (2). (5) follows from Subsection 3.5 in [38].

5.2. Map $\bullet^\dagger : \mathfrak{Id}(\mathcal{W}) \to \mathfrak{Id}(\mathcal{U})$.

By definition, $\bullet^\dagger$ is the composition

$$\mathfrak{Id}(\mathcal{W}) \to \mathfrak{Id}_\hbar(\mathcal{W}_\hbar) \to \mathfrak{Id}_\hbar(\mathcal{W}_\hbar^\wedge) \to \mathfrak{Id}_\hbar(\mathcal{U}_\hbar^\wedge) \to \mathfrak{Id}_\hbar(\mathcal{U}_\hbar) \to \mathfrak{Id}(\mathcal{U}),$$

where all maps except $\mathfrak{Id}_\hbar(\mathcal{U}_\hbar^\wedge) \to \mathfrak{Id}_\hbar(\mathcal{U}_\hbar)$ are the inverses of the corresponding maps in (5.1). The map $\mathfrak{Id}_\hbar(\mathcal{U}_\hbar^\wedge) \to \mathfrak{Id}_\hbar(\mathcal{U}_\hbar)$ sends $\mathcal{J}'_\hbar \in \mathfrak{Id}_\hbar(\mathcal{U}_\hbar^\wedge)$ to $\mathcal{J}'_\hbar \cap \mathcal{U}_\hbar$.

Let us list some properties of the map $\mathcal{I} \mapsto \mathcal{I}^\dagger : \mathfrak{Id}(\mathcal{W}) \to \mathfrak{Id}(\mathcal{U})$.

Proposition 5.2. *(1) Let N be a $\mathcal{W}$-module. Then $\operatorname{Ann}_{\mathcal{W}}(N)^\dagger = \operatorname{Ann}_{\mathcal{U}}(\mathrm{Sk}(N))$, where* Sk *denotes the Skryabin functor, see Subsection 4.1.*

(2) Let N be a $\mathcal{W}$-module from the category $\mathcal{O}(\theta)$, see Subsection 4.2. Then $\operatorname{Ann}_{\mathcal{W}}(N)^\dagger = \operatorname{Ann}_U(\mathcal{K}(N))$, where $\mathcal{K}$ is the functor from Theorem 4.3.

(3) Let $\mathcal{I}$ be an ideal of finite codimension in $\mathcal{W}$. If $\mathcal{I}$ is prime (resp., completely prime, primitive), then so is $\mathcal{I}^\dagger$.

(4) $V(\mathcal{U}/\mathcal{I}^\dagger) = \overline{\mathbb{O}}$ if and only if $\mathcal{I}$ is of finite codimension.

(5) Recall that the center $\mathcal{Z}$ of $\mathcal{U}$ is identified with the center of $\mathcal{W}$. Under this identification for any $\mathcal{I} \in \mathfrak{Id}(\mathcal{W})$ we have $\mathcal{I} \cap \mathcal{Z} = \mathcal{I}^\dagger \cap \mathcal{Z}$.

(6) The map $\mathcal{I} \mapsto \mathcal{I}^\dagger$ is Q-invariant.

Recall that an ideal I in an associative unital algebra A is called prime (resp., completely prime) if a or b lies in A whenever $aAb \subset I$ (resp., $ab \in I$). An ideal I is said to be primitive if it is the annihilator of an irreducible A-module.

(1) is assertion (ii) of [37], Theorem 1.2.2. (2) is a part of [39], Theorem 4.1. (3) stems from [37], Theorem 1.2.2. The "if" part of (4) follows from (1) and [51], Theorem 3.1. The "only if" part follows from the inclusion $(\mathcal{I}^\dagger)_\dagger \subset \mathcal{I}$ that is a direct consequence of our constructions. (5) is assertion (iii) of [37], Theorem 1.2.2. (6) follows directly from the construction.

5.3. Classification of finite dimensional irreducible $\mathcal{W}$-modules. This subsection is perhaps the most important part of the notes. Here we explain known results about the classification of finite dimensional irreducible $\mathcal{W}$-modules. We have two results here, both are due to the author, [38],[40]. Both relate the set $\mathrm{Irr}_{fin}(\mathcal{W})$ of (isomorphism classes of) finite dimensional irreducible $\mathcal{W}$-modules to the set $\mathrm{Prim}_{\mathbb{O}}(\mathcal{U})$ consisting of all primitive ideals $\mathcal{J} \subset \mathcal{U}$ with $\mathrm{V}(\mathcal{U}/\mathcal{J}) = \overline{\mathbb{O}}$.

The first result was conjectured by Premet (private communication). To state it we notice that the set $\mathrm{Irr}_{fin}(\mathcal{W})$ is canonically identified with the set $\mathrm{Prim}_{fin}(\mathcal{W})$ of maximal (=primitive) ideals of finite codimension in $\mathcal{W}$ (via taking the annihilator). Thanks to assertions (3),(4) of Proposition 5.2, we see that $N \mapsto \mathrm{Ann}_{\mathcal{W}}(N)^\dagger$ is a map $\mathrm{Irr}_{fin}(\mathcal{W}) \to \mathrm{Prim}_{\mathbb{O}}(\mathcal{U})$. The group Q acts on $\mathrm{Irr}_{fin}(\mathcal{W})$. The connected component Q° of Q acts trivially because the corresponding action of $\mathfrak{q}$ on $\mathcal{W}$ is by inner derivations. So the Q-action on $\mathrm{Irr}_{fin}(\mathcal{W})$ descends to that of the component group $C(e) = Q/Q^\circ$. By assertion (6) of Proposition 5.2, the map $\mathrm{Irr}_{fin}(\mathcal{W}) \to \mathrm{Prim}_{\mathbb{O}}(\mathcal{U})$ is $C(e)$-invariant.

Conjecture 5.3 (Premet). The map $N \mapsto \mathrm{Ann}(N)^\dagger : \mathrm{Irr}_{fin}(\mathcal{W}) \to \mathrm{Prim}_{\mathbb{O}}(\mathcal{U})$ is surjective and any of its fibers is a single $C(e)$-orbit.

In [52] Premet proved that any $\mathcal{J} \in \mathrm{Prim}_{\mathbb{O}}(\mathcal{U})$ with *rational* central character lies in the image. In full generality the surjectivity part was first proved in [37], Theorem 1.2.2. Later alternative proofs were found in [26],[53]. The description of fibers is more subtle. It was obtained in [38]. It is a corollary of the following theorem.

Theorem 5.4 ([38], Theorem 1.2.2). *Let $\mathcal{I}$ be a Q-*stable *ideal of finite codimension in $\mathcal{W}$. Then $(\mathcal{I}^\dagger)_\dagger = \mathcal{I}$.*

The second result is stated in terms of the category $\mathcal{O}(\theta)$. Let $\theta, \mathfrak{l} = \mathfrak{g}_0, \mathcal{W}^0$ have the same meaning as in Subsection 4.2. Choose a Cartan subalgebra $\mathfrak{h} \subset \mathfrak{l}$ and a system of simple roots $\Pi \subset \mathfrak{h}^*$ as in the discussion preceding Theorem 4.4.

Let us introduce some more notation. For $\lambda \in \mathfrak{h}^*$ let $L_0(\lambda)$ stand for the irreducible $\mathfrak{g}_0$-module with highest weight λ. Set $J(\lambda) := \mathrm{Ann}_{\mathcal{U}}(L(\lambda)), J_0(\lambda) := \mathrm{Ann}_{U(\mathfrak{g}_0)}(L_0(\lambda))$. According to Duflo, [19], any primitive ideal in $\mathcal{U}$ (resp., in $U(\mathfrak{g}_0)$) has the form $J(\lambda)$ (resp., $J_0(\lambda)$) for some (in general, non-unique) $\lambda \in \mathfrak{h}^*$.

Proposition 5.5. *[[40], Theorem 5.1.1] Let N_0 be an irreducible finite dimensional $\mathcal{W}^0$-module. If $\mathrm{Ann}_{\mathcal{W}^0}(N_0)^\dagger = J_0(\lambda)$ for some λ, then $\mathrm{Ann}_{\mathcal{W}}(L^\theta(N_0))^\dagger = J(\lambda)$. In particular, $L^\theta(N_0)$ is finite dimensional if and only if $\mathrm{V}(\mathcal{U}/J(\lambda)) = \overline{\mathbb{O}}$.*

5.4. Remarks. In the representation theory of $\mathcal{U}$ there are many results on the computation of $\mathrm{V}(\mathcal{U}/J(\lambda))$ and on the description of $\mathrm{Prim}_{\mathbb{O}}(\mathcal{U})$. They are due to Joseph, Barbasch-Vogan and others, see, for example, [2],[3],[29]. In particular, it is known that $\mathrm{Prim}_{\mathbb{O}}(\mathcal{U})$ is always non-empty.

Next, we remark that the maps between the sets of ideals upgrade to functors between the categories of Harish-Chandra bimodules, see [26],[38]. The study of these functors is supposed to help to obtain the complete description of $\mathrm{Irr}_{fin}(\mathcal{W})$ itself (not just of the quotient $\mathrm{Irr}_{fin}(\mathcal{W})/C(e)$).

6. One-dimensional $\mathcal{W}$-modules

6.1. Motivation. The following conjecture was made by Premet.

Conjecture 6.1 ([51], Conjecture 3.1). Any W-algebra has a one-dimensional representation (equivalently, a two-sided ideal of codimension 1).

At the moment when this text is being written the conjecture is known to be true with exception of several cases in type E_8, where it is still open.

The reason why Conjecture 6.1 is important is that it implies affirmative answers to some old questions in representation theory of universal enveloping algebras:

(A) the question of Humphreys on the existence of a small non-restricted representation for semisimple Lie algebras in characteristic p.

(B) the existence of a completely prime primitive ideal with given associated variety (this question traces back, at least, to Dixmier)

The proof that Conjecture 6.1 implies (A) for $p \gg 0$ is obtained in [53], Theorem 1.4.

The claim that Conjecture 6.1 implies (B) follows from Proposition 5.2: if $\mathcal{I} \subset \mathcal{W}$ has codimension 1, then $\mathcal{I}^\dagger$ is completely prime and $\mathrm{V}(\mathcal{U}/\mathcal{I}^\dagger) = \overline{\mathbb{O}}$.

In fact, the implication in the previous paragraph was obtained earlier by Moeglin, [46],[47]. She considered primitive ideals in $\mathcal{U}$ admitting a *Whittaker model*. Using the techniques of [25], one can show that a Whittaker model in the sense of Moeglin is precisely the image of a one-dimensional $\mathcal{W}$-module under the Skryabin equivalence.

Actually, Moeglin obtained a stronger result: that any ideal admitting a Whittaker model gives rise to a unique quantization (in an appropriate sense) of a suitable covering of $\mathbb{O}$, see [47] for details.

6.2. Classical algebras. It turns out that Conjecture 6.1 holds for all classical simple Lie algebras. This was proved in [37], Theorem 1.2.3, (1). Let us describe the idea of the proof.

We need to show that there is an ideal of codimension 1 in $\mathcal{W}$. Thanks to Proposition 5.1, this is the case when there is $\mathcal{J} \in \mathfrak{Id}(\mathcal{U})$ such that $\overline{\mathbb{O}}$ is

an irreducible component of $\mathrm{V}(\mathcal{U}/\mathcal{J})$ and the multiplicity of $\mathcal{U}/\mathcal{J}$ on $\overline{\mathbb{O}}$ is 1 (this implication also was proved by Moeglin using the language of Whittaker models, see [47]).

Let G be one of the classical groups $\mathrm{SL}_n(\mathbb{K})$, $\mathrm{O}_n(\mathbb{K}), \mathrm{Sp}_{2n}(\mathbb{K})$ (depending on $\mathfrak{g}$). We emphasize that for $\mathfrak{g} = \mathfrak{so}_n$ we need a disconnected group. It turns out that there is an ideal $\mathcal{J}$ in $\mathcal{U}$ such that $\mathrm{gr}\,\mathcal{U}/\mathcal{J} = \mathbb{K}[\overline{Ge}]$, where gr is taken with respect to the filtration on $\mathcal{U}/\mathcal{J}$ induced from the PBW filtration on $\mathcal{U}$. Such an ideal is obtained by the quantization of the Kraft-Procesi construction of $\overline{Ge}$ via a Hamiltonian reduction of a vector space, see [34],[35][1].

In type A more can be said. Form the quotient $\mathcal{W}^{ab}$ of $\mathcal{W}$ by the relations $[a, b], a, b \in \mathcal{W}$. The one-dimensional $\mathcal{W}$-modules are parametrized by points of $\mathrm{Spec}(\mathcal{W}^{ab})$. In [53], Subsection 3.8, Premet proved that for $\mathfrak{g} = \mathfrak{sl}_n$ the algebra $\mathcal{W}^{ab}$ is the polynomial algebra in $d-1$ variables, where d is the maximal size of a Jordan block of e. Premet's proof is based on the Brundan-Kleshchev presentation of $\mathcal{W}$, see Subsection 7.1 for details.

6.3. Parabolic induction. It is easy to prove Conjecture 6.1 when e is even. Indeed, as we have seen in Subsection 2.2, the algebra $\mathcal{W}$ for even e can be embedded into $U(\mathfrak{g}(0))$, see Subsection 2.2. Then we can take any 1-dimensional representation of $U(\mathfrak{g}(0))$ and restrict it to $\mathcal{W}$.

Premet, [53], observed that a similar result holds in a much more general setting. In the theory of nilpotent elements in semisimple Lie algebras there is a construction called the *Lusztig-Spaltenstein induction*. It was introduced in [41], for a review see, for example, [44]. Namely, let $\underline{\mathfrak{g}} \subset \mathfrak{g}$ be a Levi subalgebra and $\underline{\mathbb{O}} \subset \underline{\mathfrak{g}}$ be a nilpotent orbit. The Lusztig-Spaltenstein induction produces a nilpotent orbit $\mathbb{O} \subset \mathfrak{g}$ from the pair $(\underline{\mathfrak{g}}, \underline{\mathbb{O}})$. We say that $\mathbb{O}$ is induced from $(\underline{\mathfrak{g}}, \underline{\mathbb{O}})$. If e is even, then $\mathbb{O}$ is induced from $(\mathfrak{g}(0), \{0\})$. If $\mathbb{O}$ cannot be induced from a nilpotent orbit in proper Levi subalgebra, $\mathbb{O}$ is called *rigid*.

Theorem 6.2 (Premet, [53]). *Let $\mathbb{O}$ be induced from $(\underline{\mathfrak{g}}, \underline{\mathbb{O}})$. If the algebra $\underline{\mathcal{W}} := U(\underline{\mathfrak{g}}, \underline{e})$, where $\underline{e} \in \underline{\mathbb{O}}$, has a one-dimensional representation, then $\mathcal{W}$ does.*

Premet's proof of Theorem 6.2 is based on the reduction to positive characteristic. Another proof, close in spirit to that for even elements, was found by the author in [40]. Namely, under the assumptions of Theorem 6.2, there is an embedding of $\mathcal{W}$ into a certain *completion* of $\underline{\mathcal{W}}$. The latter acts on all finite dimensional $\underline{\mathcal{W}}$-modules. So a one-dimensional $\mathcal{W}$-module again can be obtained by restriction.

6.4. 1-dimensional representations via category $\mathcal{O}(\theta)$. In this subsection we will explain how to apply the category $\mathcal{O}(\theta)$ to the study of one-dimensional representations of $\mathcal{W}$, see [40].

[1] After [37] was already published I learned that the construction of $\mathcal{J}$ used there (and explained above) was discovered before by R. Brylinski,[15].

We use the notation from Subsection 4.2. Let us impose the following condition on a nilpotent element e:

(*) the algebra $\mathfrak{q}$ is semisimple.

It turns out that this condition is satisfied for all rigid nilpotent elements. A proof based on the classification of such elements can be found in [40], Subsection 5.2. It would be very interesting to find a conceptual proof.

Let N^0 be a finite dimensional $\mathcal{W}^0$-module. We want a criterium for $L^\theta(N^0)$ to be 1-dimensional. Since $N^0 \hookrightarrow L^\theta(N^0)$, of course, $\dim L^\theta(N^0) = 1$ implies $\dim N^0 = 1$.

The following result follows from Theorem 5.2.1 in [40].

Theorem 6.3. *Suppose the condition (*) holds. Let N^0 be a 1-dimensional $\mathcal{W}^0$-module. Then the following conditions are equivalent:*

1. $\dim L^\theta(N^0) = 1$.
2. $\mathfrak{t} \subset \mathcal{W}^0$ *acts on* N^0 *by* δ, *where* δ *is given by (4.1).*

When e is of principal Levi type (which is true for all but 2 rigid nilpotent elements in exceptional Lie algebras), then any irreducible $\mathcal{W}^0$-module is 1-dimensional (recall that $\mathcal{W}^0$ is just the center of $U(\mathfrak{l})$).

Combining Theorem 6.3, Proposition 5.5, and assertion (4) of Proposition 5.2 one obtains a criterium for an ideal $\mathcal{J} \subset \mathcal{U}$ to have the form $\mathcal{I}^\dagger$ with $\dim \mathcal{W}/\mathcal{I} = 1$. More precisely, we have the following result, [40], Subsection 5.3.

Corollary 6.4. *Suppose $\mathfrak{q}$ satisfies (*). Let $\theta, \mathfrak{h}, \Pi$ be chosen as in the discussion preceding Theorem 4.4. Let $\mathbb{O}_0$ denote the adjoint orbit of e in $\mathfrak{l}$.*

1. *Let $\lambda \in \mathfrak{h}^*$ satisfy the following four conditions:*
   - *(A) The associated variety of $U(\mathfrak{l})/J_0(\lambda)$ in $\mathfrak{g}_0^*$ is $\overline{\mathbb{O}_0}$.*
   - *(B) $\dim \mathrm{V}(\mathcal{U}/J(\lambda)) \leqslant \dim \mathbb{O}$.*
   - *(C) $\lambda - \delta$ vanishes on $\mathfrak{t}$.*
   - *(D) $J_0(\lambda)$ corresponds to an ideal of codimension 1 in $\mathcal{W}^0$.*

   Then $J(\lambda) = \mathcal{I}^\dagger$ for some ideal $\mathcal{I} \subset \mathcal{W}$ of codimension 1.
2. *For any ideal $\mathcal{I} \subset \mathcal{W}$ of codimension 1 there is $\lambda \in \mathfrak{h}^*$ satisfying (A)-(D) and such that $J(\lambda) = \mathcal{I}^\dagger$.*

When e is principal in $\mathfrak{l}$ the condition (A) means that λ is antidominant for $\mathfrak{l}$, while the condition (D) becomes vacuous.

6.5. Exceptional algebras. Let us summarize what is known about Conjecture 6.1 for exceptional Lie algebras. As Premet checked in [51], $\mathcal{W}$ has a one-dimensional module provided e is a minimal nilpotent element (in an arbitrary simple Lie algebra). His approach was based on an analysis of generators and relations for $\mathcal{W}$ that are not very difficult for minimal nilpotents. Recently Goodwin, Röhrle and Ubly, [28], extended Premet's approach to all rigid nilpotents in G_2, F_4, E_6, E_7 and some rigid nilpotents in E_8. The result is that for all nilpotent elements they considered a one-dimensional $\mathcal{W}$-module does exist. They used the GAP program to analyze the relations. "Large" nilpotent elements in E_8 remain to complicated computationally. Maybe, one can deduce Conjecture 6.1 for E_8 from Corollary 6.4.

7. Type *A*

This section is devoted to results concerning W-algebras for $\mathfrak{g} = \mathfrak{sl}_N$ (or $\mathfrak{g} = \mathfrak{gl}_N$). In Subsection 7.1 we very briefly sketch a relation between W-algebras and Yangians. In Subsection 7.2 we mention some other results: the higher level Schur-Weyl duality of Brundan and Kleshchev and the Gelfand-Kirillov conjecture for W-algebras proved by Futorny, Molev and Ovsienko.

7.1. W-algebras vs Yangians. In this subsection we will briefly explain a relationship between W-algebras for $\mathfrak{g} = \mathfrak{gl}_N$ and certain inifinite dimensional algebras called *shifted Yangians*. A shifted Yangian is a certain generalization of the usual Yangian for $\mathfrak{gl}_n$. For a comprehensive treatment of Yangians and related algebras the reader is referred to Molev's book [48]. A relation between Yangians and W-algebras was first observed by Ragoucy and Sorba in [56] and then generalized to shifted Yangians by Brundan and Kleshchev, [12].

The Yangian $Y(\mathfrak{gl}_n)$ can be defined as the algebra generated by elements $t_{ij}^{(r)}, i, j = 1, \dots, n, r \in \mathbb{N}$, subject to the relations

$$\left[t_{ij}^{(r+1)}, t_{kl}^{(s)}\right] - \left[t_{ij}^{(r)}, t_{kl}^{(s+1)}\right] = t_{kj}^{(r)} t_{il}^{(s)} - t_{kj}^{(s)} t_{il}^{(r)}.$$

However, the generators $t_{ij}^{(r)}$ are not convenient to establish a relation between the Yangians and W-algebras. In [11] Brundan and Kleshchev found a new presentation of $Y(\mathfrak{gl}_n)$. Generalizing this presentation they introduced shifted Yangians in [12].

A shifted Yangian $Y_n(\sigma)$ depends on a positive integer n and some *shift matrix* σ. By definition, $\sigma = (s_{ij})_{i,j=1}^n$ is a shift matrix if s_{ij} is a nonnegative integer ("shift") with $s_{ij} + s_{jk} = s_{ik}$ whenever $|i-j| + |j-k| = |i-j|$. By definition, the algebra $Y_n(\sigma)$ is given by generators

$$D_i^{(r)}(1 \leqslant i \leqslant n, r > 0), E_i^{(r)}(1 \leqslant i < n, r > s_{i,i+1}), F_i^{(r)}(1 \leqslant i < n, r > s_{i+1,i})$$

subject to certain explicit relations (see [12], (2.4)-(2.15)). The shifted Yangian coincides with the usual one when $\sigma = 0$. For $l > s_{1,n} + s_{n,1}$ define the quotient (the truncated shifted Yangian of level l) $Y_{n,l}(\sigma)$ of $Y_n(\sigma)$ by the two-sided ideal generated by $D_1^{(r)}, r > p_1 := l - s_{1,n} - s_{n,1}$.

To establish a relationship between shifted Yangians and W-algebras fix a positive integer n, pick a Young diagram $\lambda = (\lambda_1, \ldots, \lambda_n), \lambda_1 \geqslant \ldots \geqslant \lambda_n \geqslant 0$ (one can also work with more general diagrams called *pyramids*, see [12], §7 for details), and set $l := \lambda_1$. Then to n and λ one can assign the *shift matrix* $\sigma^\lambda = (s_{ij})_{i,j=1}^n$ by setting $s_{ij} := 0$ for $i \geqslant j$ and $s_{ij} := \lambda_{n+1-j} - \lambda_{n+1-i}$ for $i < j$. In particular, for the Young diagram of shape $n \times l$, we get $\sigma = 0$.

To λ one assigns a nilpotent element $e_\lambda \in \mathfrak{gl}_N$, where $N := \sum_{i=1}^n \lambda_i$, in the usual way ($\lambda_i$ are the sizes of the Jordan blocks of e_λ).

Theorem 7.1 ([12], Theorem 10.1). $U(\mathfrak{gl}_N, e_\lambda) \cong Y_{n,l}(\sigma^\lambda)$.

In [13] Brundan and Kleshchev used this theorem to study the representation theory of $U(\mathfrak{gl}_N, e_\lambda)$. In particular, they obtained a classification of finite dimensional irreducible $U(\mathfrak{gl}_N, e)$-modules (which also follows from Proposition 5.5 thanks to Joseph's computation of $\mathrm{V}(U(\mathfrak{gl}_N)/J(\lambda))$, see [29]; we remark that any nilpotent element in $\mathfrak{gl}_N$ is of principal Levi type).

There is a generalization of the results explained above in this subsection to other classical Lie algebras first observed by Ragoucy, [55] and worked out in more detail by J. Brown, [7],[8]. Namely, for orthogonal and symplectic algebras there are analogs of $Y(\mathfrak{gl}_n)$ called *twisted* Yangians. Theorem 7.1 generalizes to twisted Yangians. It is interesting that, similarly to $Y(\mathfrak{gl}_n)$-case, nilpotent elements arising in this generalization again correspond to partitions with all parts equal. It is unclear whether there is a reasonable shifted version of the twisted Yangians that is related to the W-algebras constructed from arbitrary nilpotent elements.

7.2. Other results. W-algebras in type A enjoy some other interesting properties.

For example, in [14] Brundan and Kleshchev obtained a very nice result: a "higher level" generalization of the classical Schur-Weyl duality. Recall that the classical Schur-Weyl duality relates between polynomial representations of $\mathrm{GL}_N(\mathbb{K})$ and representations of the symmetric group S_d in d letters. The Brundan-Kleshchev generalization relates modules over the cyclotomic degenerate Hecke algebra $H_d(\lambda)$ corresponding to a partition λ of N (this algebra is a higher level generalization of S_d) and modules over the W-algebra $U(\mathfrak{gl}_N, e_\lambda)$. For details the reader is referred to [14] or to the review [60] by Wang.

Another result we would like to mention is an analog of the Gelfand-Kirillov conjecture for W-algebras proved in [23].

For a Noetherian domain A let $Q(A)$ denote its skew-field of fractions. Gelfand and Kirillov, [24], conjectured that for any finite dimensional algebraic Lie algebra $\mathfrak{a}$ the skew-field $Q(U(\mathfrak{a}))$ is isomorphic to $Q(\mathbf{A}_l(F_d))$, where F_d is

a purely transcendental extension of $\mathbb{K}$ of some degree d and $\mathbf{A}_l(F_d)$ stands for the Weyl algebra of a $2l$-dimensional symplectic vector space over F_d. In [24] the conjecture was verified for $\mathfrak{g} = \mathfrak{sl}_n$[2]. In [23] Futorny, Molev and Ovsienko proved that the straightforward analog of the Gelfand-Kirillov conjecture holds for $U(\mathfrak{gl}_n, e)$ (and for $U(\mathfrak{sl}_n, e)$) for an arbitrary nilpotent element $e \in \mathfrak{sl}_n$.

References

[1] E. Backelin, *Representation of the category $\mathcal{O}$ in Whittaker categories*, IMRN, 4(1997), 153–172.

[2] D. Barbasch, D. Vogan, *Primitive ideals and orbital integrals in complex classical groups*, Math. Ann. 259(1982), 153–199.

[3] D. Barbasch, D. Vogan, *Primitive ideals and orbital integrals in complex exceptional groups*, J. Algebra 80(1983), 350–382.

[4] A. Beilinson, J. Bernstein, *Localization de $\mathfrak{g}$-modules*, C. R. Acad. Sci. Paris 292 (1981), no. 1, 15–18.

[5] I. Bernstein, I. Gelfand, S. Gelfand, *A category of $\mathfrak{g}$-modules*, Funct. Anal. Appl. 10(1976), 87–92.

[6] J. de Boer, T. Tjin, *Quantization and representation theory of finite W-algebras*, Comm. Math. Phys. 158(1993), 485–516.

[7] J. Brown, *Twisted Yangians and finite W-algebras*, Transform. Groups 14 (2009), 87–114.

[8] J. Brown, *Representation theory of rectangular finite W-algebras*, arXiv: 1003.2179.

[9] J. Brundan, S. Goodwin, *Good gradings polytopes*, Proc. London Math. Soc. 94(2007), 155–180.

[10] J. Brundan, S. Goodwin, A. Kleshchev, *Highest weight theory for finite W-algebras*, IMRN 2008, no. 15, Art. ID rnn051.

[11] J. Brundan, A. Kleshchev, *Parabolic presentation of the Yangian $Y(\mathfrak{gl}_n)$*, Comm. Math. Phys. 254(2005), 191–220.

[12] J. Brundan, A. Kleshchev, *Shifted Yangians and finite W-algebras*, Adv. Math. 200(2006), 136–195.

[13] J. Brundan, A. Kleshchev, *Representations of shifted Yangians and finite W-algebras*, Mem. Amer. Math. Soc. 196 (2008), 107 pp.

[14] J. Brundan, A. Kleshchev, *Schur-Weyl duality for higher levels*, Selecta Math., 14(2008), 1–57.

[15] R. Brylinski, *Dixmier algebras for classical complex nilpotent orbits via Kraft-Procesi models. I*, Prog. Math. 213, Birkhäuser, Boston, 49–67.

[2] Recently Premet proved in [54] that the Gelfand-Kirillov conjecture does not hold for $\mathfrak{a}$ of type $B_n (n \geqslant 3), D_n (n \geqslant 4), E_6, E_7, E_8, F_4$.

[16] A. D'Andrea, C. De Concini, A. De Sole, R. Heluani and V. Kac, *Three equivalent definitions of finite W-algebras*, Appendix to [17].

[17] A. De Sole, V. Kac, *Finite vs affine W-algebras*, Japan. J. Math, 1(2006), 137–261.

[18] C. Dodd, K. Kremnizer, *A Localization Theorem for Finite W-algebras*, arXiv:0911.2210.

[19] M. Duflo, *Sur la classification des idéaux primitifs dans l'algèbre envellopante d'une algèbre de Lie semi-simple*, Ann. Math. 105(1977), 107–120.

[20] A. Elashvili, V. Kac, *Classification of good gradings of simple Lie algebras*, in: "Lie groups and invariant theory" (E.B. Vinberg ed.), Amer. Math. Soc. Transl. ser. 2, 213(2005), 85–104.

[21] B. Fedosov, *A simple geometrical construction of deformation quantization*, J. Diff. Geom. 40(1994), 213–238.

[22] B. Fedosov, *Deformation quantization and index theory*, in Mathematical Topics 9, Akademie Verlag, 1996.

[23] V. Futorny, A. Molev, S. Ovsienko, *Gelfand-Kirillov conjecture and Gelfand-Tsetlin modules for finite W-algebras*, Adv. Math. 223(2010), 773–796.

[24] I. Gelfand, A. Kirillov, *Sur les corps lés aux algèbres enveloppantes des algèbres de Lie*, Publ. IHES, 31(1966), 5–19.

[25] W.L. Gan, V. Ginzburg, *Quantization of Slodowy slices*, IMRN, 5(2002), 243–255.

[26] V. Ginzburg, *Harish-Chandra bimodules for quantized Slodowy slices*, Repres. Theory 13(2009), 236–271.

[27] S. Goodwin, *Translation for finite W-algebras*, arXiv:0908.2739.

[28] S. Goodwin, G. Röhrle, G. Ubly, *On 1-dimensional representations of finite W-algebras associated to simple Lie algebras of exceptional type*, arXiv:0905.3714.

[29] A. Joseph, *Sur la classification des idéaux primitifs dans l'algebre envellopante de $\mathfrak{sl}(n+1,\mathbb{C})$*, C.R. Acad. Sci. Paris Sér A–B, 287(1978), N5, A303–306.

[30] N. Kawanaka, *Generalized Gelfand-Graev representations and Ennola duality*, In "Algebraic Groups and Related Topics", Advanced Studies in Pure Mathematics 6(1985), North-Holland, p. 175–206.

[31] B. Kostant, *Lie group representations on polynomial rings*, Amer. J. Math. 85(1963), 327–404.

[32] B. Kostant, *On Whittaker vectors and representation theory*, Invent. Math. 48(1978), 101–184.

[33] B. Kostant, *The Solution to a Generalized Toda Lattice and Representation Theory*, Adv. in Math., 34(1979), 195–338.

[34] H. Kraft, C. Procesi, *Closures of conjugacy classes of matrices are normal*, Invent. Math. 53(1979), 227–247.

[35] H. Kraft, C. Procesi, *On the geometry of conjugacy classes in classical groups*, Comment. Math. Helv. 57(1982), 539–602.

[36] I.V. Losev, *Symplectic slices for reductive groups*, Mat. Sbornik 197(2006), N2, 75–86 (in Russian). English translation in: Sbornik Math. 197(2006), N2, 213–224.

[37] I.V. Losev, *Quantized symplectic actions and W-algebras*, J. Amer. Math. Soc. 23(2010), 35–59.

[38] I. Losev, *Finite dimensional representations of W-algebras*, arXiv:0807.1023.

[39] I. Losev. *On the structure of the category $\mathcal{O}$ for W-algebras*, arXiv:0812.1584.

[40] I. Losev, *1-dimensional representations and parabolic induction for W-algebras*, arXiv:0906.0157.

[41] G. Lusztig, N. Spaltenstein, *Induced unipotent classes*, J. London Math. Soc. (2), 19(1979), 41–52.

[42] T.E. Lynch, *Generalized Whittaker vectors and representation theory*, Thesis, M.I.T., 1979.

[43] E. McDowell, *On modules induced from Whittaker modules*, J. Algebra 96(1985), n.1, 161–177.

[44] W. McGovern, *The adjoint representation and the adjoint action*, Encyclopaedia of mathematical sciences, 131. Invariant theory and algebraic transformation groups, II, Springer Verlag, Berlin, 2002.

[45] D. Milicic, W. Soergel. *The composition series of modules induced from Whittaker modules*, Comment. Math. Helv. 72(1997), 503–520.

[46] C. Moeglin, *Modèles de Whittaker et idéaux primitifs complètement premiers dans les algèbres enveloppantes I*, C.R. Acad. Sci. Paris, Sér. I 303(1986), No. 17, 845–848.

[47] C. Moeglin, *Modèles de Whittaker et idéaux primitifs complètement premiers dans les algèbres enveloppantes II*, Math. Scand. 63 (1988), 5–35.

[48] A. Molev, *Yangians and classical Lie algebras*, Mathematical Surveys and Monographs, 142. AMS (2007).

[49] A. Premet, *Irreducible representations of Lie algebras of reductive groups and the Kac-Weisfeiler conjecture*, Invent. Math. 121(1995), 79–117.

[50] A. Premet, *Special transverse slices and their enveloping algebras*, Adv. Math. 170(2002), 1–55.

[51] A. Premet, *Enveloping algebras of Slodowy slices and the Joseph ideal*, J. Eur. Math. Soc, 9(2007), N3, 487–543.

[52] A. Premet, *Primitive ideals, non-restricted representations and finite W-algebras.* Moscow Math. J. 7(2007), 743–762.

[53] A. Premet, *Commutative quotients of finite W-algebras*, arXiv:0809.0663. Accepted by Adv. Math.

[54] A. Premet, *Modular Lie algebras and Gelfand-Kirillov conjecture*, arXiv:0907.2500.

[55] E. Ragoucy, *Twisted Yangians and folded W-algebras*, Internat. J. Modern. Phys. A 16 (2001), 2411–2433

[56] E. Ragoucy, P. Sorba, *Yangian realizations from finite W-algebras*, Comm. Math. Phys. 203(1999), 551–572.

[57] S. Skryabin, An appendix to [50].

[58] P. Slodowy, *Simple singularities and simple algebraic groups*, Lect. Notes Math., v.815. Springer, Berlin/Heidelberg/New York, 1980.

[59] K. de Vos, P. van Driel, *Kazhdan-Lusztig conjecture for finite W-algebras*, Lett. Math. Phys. 35(1996), 333–344.

[60] W. Wang, *Nilpotent orbits and W-algebras*, arXiv:0912.0689.

[61] B. Webster, *Singular blocks of parabolic category O and finite W-algebras*, arXiv:0909.1860.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Dynamics on Geometrically Finite Hyperbolic Manifolds with Applications to Apollonian Circle Packings and Beyond

Hee Oh*

Abstract

We present recent results on counting and distribution of circles in a given circle packing invariant under a geometrically finite Kleinian group and discuss how the dynamics of flows on geometrically finite hyperbolic 3 manifolds are related. Our results apply to Apollonian circle packings, Sierpinski curves, Schottky dances, etc.

Mathematics Subject Classification (2010). Primary 37A17, Secondary 37A40

Keywords. Circles, Apollonian circle packings, geometrically finite groups, Patterson-Sullivan density

1. Introduction

Let G be a connected semisimple Lie group and $\Gamma < G$ a discrete subgroup with finite co-volume. Dynamics of flows on the homogeneous space $\Gamma\backslash G$ have been studied intensively over the last several decades and brought many surprising applications in various fields notably including analytic number theory, arithmetic geometry and Riemmanian geometry (see [45], [58], [12], [18], [32], [78], [41], [74], [16], [21], [22], [76], [15], [49], [33], [75], [25], [27], [26], [66], etc.) The assumption that the volume of $\Gamma\backslash G$ is finite is crucial in most developments in the ergodic theory for flows on $\Gamma\backslash G$, as many basic ergodic theorems fail in the setting of an infinite measure space. It is unclear what kind of measure theoretic

*Mathematics department, Brown university, Providence, RI, U.S.A., and Korea Institute for Advanced Study, Seoul, Korea. E-mail: heeoh@math.brown.edu.

and topological rigidity for flows on $\Gamma\backslash G$ can be expected for a general discrete subgroup Γ.

In this article we consider the situation when G is the isometry group of the real hyperbolic space $\mathbb{H}^n$, $n \geq 2$, and $\Gamma < G$ is a geometrically finite discrete subgroup. In such cases we have a rich theory of the Patterson-Sullivan density and the structure of a fundamental domain for Γ in $\mathbb{H}^n$ is well understood. Using these we obtain certain equidistribution results for specific flows on the unit tangent bundle $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n)$ and apply them to prove results on counting and equidistribution for circles in a given circle packing of the plane (and also of the sphere) invariant under geometrically finite groups.

There are numerous natural questions which arise from the analogy with the finite volume cases and most of them are unsolved. We address some of them in the last section. We remark that an article by Sarig [61] discusses related issues but for geometrically *infinite* surfaces.

Acknowledgement: I would like to thank Peter Sarnak for introducing Apollonian circle packings to me and for the encouragement to work on this project. I am grateful to Curt McMullen for showing me the picture of Sierpinski curve which led me to think about more general circle packings beyond Apollonian ones, as well as for many valuable discussions. I thank my collaborators Nimish Shah and Alex Kontorovich for the joint work. I also thank Marc Burger and Gregory Margulis for carefully reading an earlier draft and making many helpful comments. Finally I thank my family for their love and support always.

2. Preliminaries

We review some of basic definitions as well as set up notations. Let G be the identity component of the isometry group of the real hyperbolic space $\mathbb{H}^n$, $n \geq 2$. Let $\Gamma < G$ be a torsion-free discrete subgroup. We denote by $\partial_\infty(\mathbb{H}^n)$ the geometric boundary of $\mathbb{H}^n$. The limit set $\Lambda(\Gamma)$ of Γ is defined to be the set of accumulation points of an orbit of Γ in $\mathbb{H}^n \cup \partial_\infty(\mathbb{H}^n)$. As Γ acts on $\mathbb{H}^n$ properly discontinuously, $\Lambda(\Gamma)$ lies in $\partial_\infty(\mathbb{H}^n)$. Its complement $\Omega(\Gamma) := \partial_\infty(\mathbb{H}^n) - \Lambda(\Gamma)$ is called the domain of discontinuity for Γ.

An element $g \in G$ is called parabolic if it fixes a unique point in $\partial_\infty(\mathbb{H}^n)$ and loxodromic if it fixes two points in $\partial_\infty(\mathbb{H}^n)$. A limit point $\xi \in \Lambda(\Gamma)$ is called a parabolic fixed point if it is fixed by a parabolic element of Γ and called a radial limit point (or a conical limit point or a point of approximation) if for some geodesic ray β tending to ξ and some point $x \in \mathbb{H}^n$, there is a sequence $\gamma_i \in \Gamma$ with $\gamma_i x \to \xi$ and $d(\gamma_i x, \beta)$ is bounded, where d denotes the hyperbolic distance. A parabolic fixed point ξ is called bounded if $\mathrm{Stab}_\Gamma(\xi)\backslash(\Lambda(\Gamma) - \{\xi\})$ is compact.

The convex core C_Γ of Γ is defined to be the minimal convex set in $\mathbb{H}^n$ mod Γ which contains all geodesics connecting any two points in $\Lambda(\Gamma)$. A discrete

subgroup Γ is called *geometrically finite* if the unit neighborhood of its convex core has finite volume and called *convex co-compact* if its convex core is compact. It is clear that a (resp. co-compact) lattice in G is geometrically finite (resp. convex co-compact). Bowditch showed [5] that Γ is geometrically finite if and only if $\Lambda(\Gamma)$ consists entirely of radial limit points and bounded parabolic fixed points. It is further equivalent to saying that Γ is finitely generated for $n = 2$, and that Γ admits a finite sided fundamental domain in $\mathbb{H}^3$ for $n = 3$. We refer to [5] for other equivalent definitions.

Γ is called *elementary* if $\Lambda(\Gamma)$ consists of at most two points, or equivalently, Γ has an abelian subgroup of finite index.

We denote by $0 \le \delta_\Gamma \le n-1$ the critical exponent of Γ, that is, the abscissa of convergence of the Poincare series of Γ:

$$\mathcal{P}_\Gamma(s) := \sum_{\gamma\in\Gamma} e^{-sd(o,\gamma o)}$$

where $o \in \mathbb{H}^n$. For a non-elementary group Γ, δ_Γ is positive and Sullivan [71] showed that for Γ geometrically finite, δ_Γ is equal to the Hausdorff dimension of the limit set $\Lambda(\Gamma)$.

For $\xi \in \partial_\infty(\mathbb{H}^n)$ and $y_1, y_2 \in \mathbb{H}^n$, the Busemann function $\beta_\xi(y_1, y_2)$ measures a signed distance between horospheres passing through y_1 and y_2 based at ξ:

$$\beta_\xi(y_1, y_2) = \lim_{t\to\infty} d(y_1, \xi_t) - d(y_2, \xi_t)$$

where ξ_t is a geodesic ray toward ξ.

For a vector u in the unit tangent bundle $\mathrm{T}^1(\mathbb{H}^n)$, we define $u^\pm \in \partial_\infty(\mathbb{H}^n)$ to be the two end points of the geodesic determined by u:

$$u^+ := \lim_{t\to\infty} g^t(u) \quad \text{and} \quad u^- := \lim_{t\to-\infty} g^t(u)$$

where $\{g^t\}$ denotes the geodesic flow.

We denote by $\pi : \mathrm{T}^1(\mathbb{H}^n) \to \mathbb{H}^n$ the canonical projection. Fixing a base point $o \in \mathbb{H}^n$, the map

$$u \mapsto (u^+, u^-, \beta_{u^-}(\pi(u), o))$$

yields a homeomorphism between $\mathrm{T}^1(\mathbb{H}^n)$ and $(\partial_\infty(\mathbb{H}^n) \times \partial_\infty(\mathbb{H}^n) - \{(\xi,\xi) : \xi \in \partial_\infty(\mathbb{H}^n)\}) \times \mathbb{R}$.

Throughout the paper we assume that Γ is non-elementary.

Patterson-Sullivan density: Generalizing the work of Patterson [55] for $n = 2$, Sullivan [71] constructed a Γ-invariant conformal density $\{\nu_x : x \in \mathbb{H}^n\}$ of dimension δ_Γ on $\Lambda(\Gamma)$. That is, each ν_x is a finite Borel measure on $\partial_\infty(\mathbb{H}^n)$

supported on $\Lambda(\Gamma)$ satisfying that for any $x, y \in \mathbb{H}^n$, $\xi \in \partial_\infty(\mathbb{H}^n)$ and $\gamma \in \Gamma$,

$$\gamma_* \nu_x = \nu_{\gamma x} \quad \text{and} \quad \frac{d\nu_y}{d\nu_x}(\xi) = e^{-\delta_\Gamma \beta_\xi(y,x)},$$

where $\gamma_* \nu_x(R) = \nu_x(\gamma^{-1}(R))$.

For Γ geometrically finite, such conformal density $\{\nu_x\}$ exists uniquely up to homothety. In fact, fixing $o \in \mathbb{H}^n$, $\{\nu_x\}$ is a constant multiple of the following family $\{\nu_{x,o}\}$ where $\nu_{x,o}$ is the weak-limit as $s \to \delta_\Gamma^+$ of the family of measures

$$\nu_{x,o}(s) := \frac{1}{\sum_{\gamma \in \Gamma} e^{-sd(o,\gamma o)}} \sum_{\gamma \in \Gamma} e^{-sd(x,\gamma o)} \delta_{\gamma o}$$

where $\delta_{\gamma o}$ denotes the Dirac measure at γo.

Consider the Laplacian Δ on $\mathbb{H}^n$. In the upper half-space coordinates $\mathbb{H}^n = \{(x_1, \cdots, x_{n-1}, y) : y > 0\}$ with the metric $\frac{\sqrt{dx_1^2 + \cdots + dx_{n-1}^2 + dy^2}}{y}$, it is given as

$$\Delta = -y^2 \left(\frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_{n-1}^2} + \frac{\partial^2}{\partial y^2} \right) + (n-2) y \frac{\partial}{\partial y}$$

(strictly speaking, this is the negative of the usual hyperbolic Laplacian). Sullivan [71] showed that

$$\phi_\Gamma(x) := |\nu_x|$$

is an eigenfunction for Δ with eigenvalue $\delta_\Gamma(n - 1 - \delta_\Gamma)$. From the Γ-invariance of the Patterson-Sullivan density $\{\nu_x\}$, ϕ_Γ is a function on $\Gamma \backslash \mathbb{H}^n$. Sullivan further showed that if Γ geometrically finite and $\delta_\Gamma > (n-1)/2$, ϕ_Γ belongs to $L^2(\Gamma \backslash \mathbb{H}^n)$ and is a unique (up to a constant multiple) positive eigenfunction with the smallest eigenvalue $\delta_\Gamma(n - 1 - \delta_\Gamma)$ (cf. [73]). Combined with a result of Yau [77], it follows that $\delta_\Gamma = n - 1$ if and only if Γ is a lattice in G.

Bowen-Margulis-Sullivan measure: Fixing the Patterson-Sullivan density $\{\nu_x\}$, the Bowen-Margulis-Sullivan measure m_Γ^{BMS} ([6], [46], [72]) is the induced measure on $\mathrm{T}^1(\Gamma \backslash \mathbb{H}^n)$ of the following Γ-invariant measure on $\mathrm{T}^1(\mathbb{H}^n)$:

$$d\tilde{m}^{\mathrm{BMS}}(u) = e^{\delta_\Gamma \beta_{u^+}(x, \pi(u))} \, e^{\delta_\Gamma \beta_{u^-}(x, \pi(u))} \, d\nu_x(u^+) d\nu_x(u^-) dt$$

where $x \in \mathbb{H}^n$.

It follows from the conformality of $\{\nu_x\}$ that this definition is independent of the choice of x. The measure m_Γ^{BMS} is invariant under the geodesic flow and is supported on the non-wandering set $\{u \in \mathrm{T}^1(\Gamma \backslash \mathbb{H}^n) : u^\pm \in \Lambda(\Gamma)\}$ of the geodesic flow. Sullivan showed that for Γ geometrically finite, the total mass $|m_\Gamma^{\mathrm{BMS}}|$ is finite and the geodesic flow is ergodic with respect to m_Γ^{BMS} [72]. This is a very important point for the ergodic theory on geometrically finite hyperbolic manifolds, since despite of the fact that the Liouville measure is infinite, we do have a finite measure on $\mathrm{T}^1(\Gamma \backslash \mathbb{H}^n)$ which is invariant and ergodic for the geodesic flow. Rudolph [60] showed that the geodesic flow is even mixing with respect to m_Γ^{BMS}.

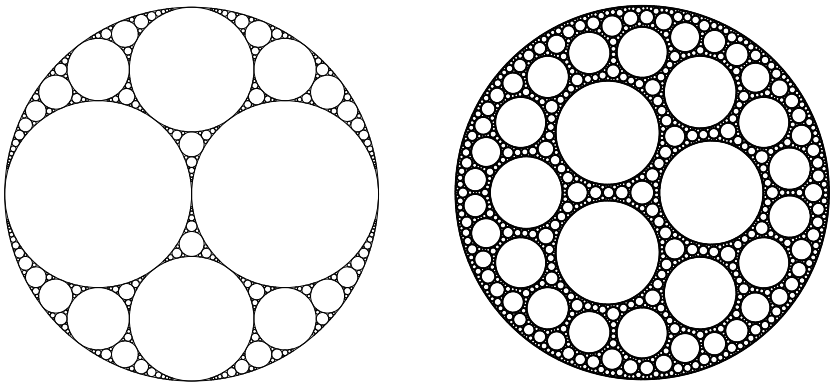


Figure 1. Apollonian circle packing and Sierpinski curve (by C. McMullen)

3. Counting and Distribution of Circles in the Plane

A circle packing in the plane $\mathbb{C}$ is simply a union of circles. As circles may intersect with each other beyond tangency points, our definition of a circle packing is more general than what is usually thought of. For a given circle packing $\mathcal{P}$ in the plane, we discuss questions on counting and distribution of small circles in $\mathcal{P}$. A natural size of a circle is measured by its radius. We will use the curvature (=the reciprocal of the radius) of a circle instead.

We suppose that $\mathcal{P}$ is infinite and that $\mathcal{P}$ is locally finite in the sense that for any $T > 0$, there are only finitely many circles of curvature at most T in any fixed bounded region of the plane. See Fig. 1, 6 and 8 for examples of locally finite packings.

For a bounded region E in the plane $\mathbb{C}$, we consider the following counting function:

$$N_T(\mathcal{P}, E) := \#\{C \in \mathcal{P} : C \cap E \neq \emptyset, \ \ \mathrm{Curv}(C) < T\}$$

where $\mathrm{Curv}(C)$ denotes the curvature of C. The local finiteness assumption is so that $N_T(\mathcal{P}, E) < \infty$ for any bounded E. We ask if there is an asymptotic for $N_T(\mathcal{P}, E)$ as T tends to infinity and what the dependence of such an asymptotic on E is, if exists.

Consider the upper half space model $\mathbb{H}^3 = \{(z, r) : z \in \mathbb{C}, r > 0\}$ with the hyperbolic metric given by $\frac{\sqrt{|dz|^2+dr^2}}{r}$. An elementary but helpful observation is that if we denote by $\hat{C} \subset \mathbb{H}^3$ the convex hull of C, that is, the northern hemisphere above C, then $N_T(\mathcal{P}, E)$ is equal to the number of hemispheres of height at most T^{-1} in $\mathbb{H}^3$ whose boundaries lie in $\mathcal{P}$ and intersect E, as the radius of a circle is same as the height of the corresponding hemisphere.

Let $\Gamma < \mathrm{PSL}_2(\mathbb{C})$ be a geometrically finite discrete subgroup and fix a Γ-invariant Patterson-Sullivan density $\{\nu_x : x \in \mathbb{H}^3\}$.

In order to present our theorem on the asymptotic of $N_T(\mathcal{P}, E)$ for $\mathcal{P}$ invariant under Γ, we introduce two new invariants associated to Γ and $\mathcal{P}$. The first one is a Borel measure on $\mathbb{C}$ which depends only on Γ.

Definition 3.1. Define a Borel measure ω_Γ on $\mathbb{C}$: for $\psi \in C_c(\mathbb{C})$,

$$\omega_\Gamma(\psi) = \int_{z\in\mathbb{C}} \psi(z) e^{\delta_\Gamma \beta_z(x, z+j)}\, d\nu_x(z)$$

where $j = (0, 1) \in \mathbb{H}^3$ and $x \in \mathbb{H}^3$. By the conformal property of $\{\nu_x\}$, this definition is independent of the choice of $x \in \mathbb{H}^3$.

Note that ω_Γ is supported on $\Lambda(\Gamma) \cap \mathbb{C}$ and in particular that $\omega_\Gamma(E) > 0$ if the interior of E intersects $\Lambda(\Gamma) \cap \mathbb{C}$ non-trivially. We compute:

$$d\omega_\Gamma = (|z|^2 + 1)^{\delta_\Gamma} d\nu_j.$$

The second one is a number in $[0, \infty]$ measuring certain size of $\mathcal{P}$:

Definition 3.2 (The Γ-skinning size of $\mathcal{P}$). For a circle packing $\mathcal{P}$ invariant under Γ, we define:

$$\mathrm{sk}_\Gamma(\mathcal{P}) := \sum_{i\in I} \int_{s\in \mathrm{Stab}_\Gamma(C_i^\dagger)\backslash C_i^\dagger} e^{\delta_\Gamma \beta_{s^+}(x, \pi(s))} d\nu_x(s^+)$$

where $x \in \mathbb{H}^3$, $\{C_i : i \in I\}$ is a set of representatives of Γ-orbits in $\mathcal{P}$ and $C_i^\dagger \subset \mathrm{T}^1(\mathbb{H}^3)$ is the set of unit normal vectors to the convex hull $\hat{C}_i$ of C_i. Again by the conformal property of $\{\nu_x\}$, the definition of $\mathrm{sk}_\Gamma(\mathcal{P})$ is independent of the choice of x and the choice of representatives $\{C_i\}$.

We remark that the value of $\mathrm{sk}_\Gamma(\mathcal{P})$ can be zero or infinite in general and we do not assume any condition on $\mathrm{Stab}_\Gamma(C_i^\dagger)$'s (they may even be trivial). By the interior of a circle C, we mean the open disk which is enclosed by C. We then have the following:

Theorem 3.3 ([51]). *Let Γ be a non-elementary geometrically finite discrete subgroup of $\mathrm{PSL}_2(\mathbb{C})$ and let $\mathcal{P} = \cup_{i\in I}\Gamma(C_i)$ be an infinite, locally finite, and Γ-invariant circle packing with finitely many Γ-orbits.*

Suppose one of the following conditions hold:

1. *Γ is convex co-compact;*
2. *all circles in $\mathcal{P}$ are mutually disjoint;*
3. *$\cup_{i\in I} C_i^\circ \subset \Omega(\Gamma)$ where C_i° denotes the interior of C_i.*

For any bounded region E of $\mathbb{C}$ whose boundary is of zero Patterson-Sullivan measure, we have

$$N_T(\mathcal{P}, E) \sim \frac{\mathrm{sk}_\Gamma(\mathcal{P})}{\delta_\Gamma \cdot |m_\Gamma^{\mathrm{BMS}}|} \cdot \omega_\Gamma(E) \cdot T^{\delta_\Gamma} \qquad \text{as } T \to \infty$$

and $0 < \mathrm{sk}_\Gamma(\mathcal{P}) < \infty$.

Remark 3.4.

1. If Γ is Zariski dense in $\mathrm{PSL}_2(\mathbb{C})$, considered as a real algebraic group, any real algebraic curve has zero Patterson-Sullivan measure [23, Cor. 1.4]. Hence the above theorem applies to any Borel subset E whose boundary is a countable union of real algebraic curves.

2. We call the complement in $\hat{\mathbb{C}}$ of the set $\cup_{i\in I}\Gamma(C_i^\circ)$ the residual set of $\mathcal{P}$. The condition (3) above is then equivalent to saying that $\Lambda(\Gamma)$ is contained in the residual set of $\mathcal{P}$.

3. If we denote by $H_\infty^-(j)$ the contracting horosphere based at ∞ in $\mathrm{T}^1(\mathbb{H}^3)$ which consists of all upward normal unit vectors on $\mathbb{C} + j = \{(z, 1) : z \in \mathbb{C}\}$, we can alternatively write the measure ω_Γ as follows:

$$\omega_\Gamma(\psi) = \int_{u\in H_\infty^-(j)} \psi(u^-) e^{\delta_\Gamma \beta_{u^-}(x,\pi(u))}\, d\nu_x(u^-)$$

   and recognize that ω_Γ is the projection of the conditional of the Bowen-Margulis-Sullivan measure $\tilde{m}^{\mathrm{BMS}}$ on the horosphere $H_\infty^-(j)$ to $\mathbb{C}$ via the map $u \mapsto u^-$. It is worthwhile to note that the hyperbolic metric on $\mathbb{C}+j$ is precisely the Euclidean metric.

4. Suppose that circles in $\mathcal{P}$ are disjoint possibly except for tangency points and that $\Lambda(\Gamma)$ is equal to the residual set of $\mathcal{P}$. If ∞ is either in $\Omega(\Gamma)$ (that is, $\mathcal{P}$ is bounded) or a parabolic fixed point for Γ, then δ_Γ is equal to the circle packing exponent $e_{\mathcal{P}}$ given by

$$e_{\mathcal{P}} = \inf\left\{ s : \sum_{C\in\mathcal{P}} r(C)^s < \infty \right\} = \sup\left\{ s : \sum_{C\in\mathcal{P}} r(C)^s = \infty \right\}$$

   where $r(C)$ denotes the radius of C [54]. This extends the earlier work of Boyd [7] on bounded Apollonian circle packings.

We discuss some concrete circle packings to which our theorem applies.

3.1. Apollonian circle packings in the plane. Apollonian circle packings are one of the most beautiful circle packings whose construction can be described in a very simple manner based on an old theorem of Apollonius (262-190 BC). It says that given three mutually tangent circles in the plane, there are exactly two circles which are tangent to all the three circles.

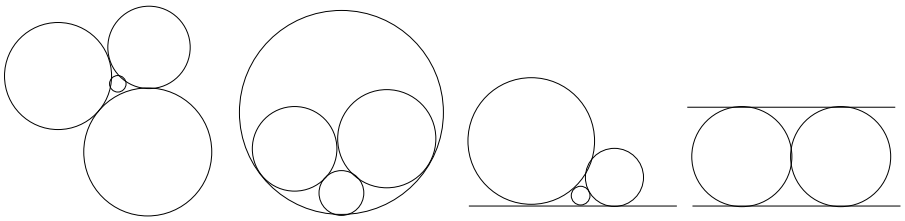


Figure 2. Possible configurations of four mutually tangent circles

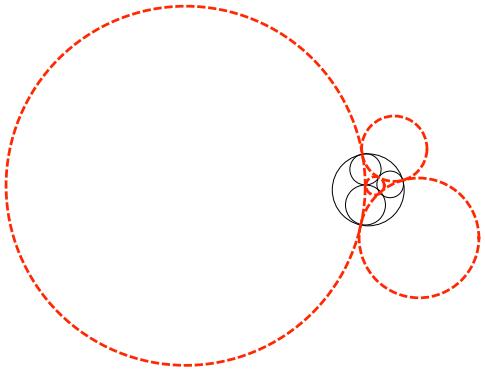


Figure 3. Dual circles

In order to construct an Apollonian circle packing, we start with four mutually tangent circles. See Fig. 2 for possible configurations. By Apollonius' theorem, there are precisely four new circles that are tangent to three of the four circles. Continuing to repeatedly add new circles tangent to three of the circles from the previous generations, we arrive at an infinite circle packing, called an Apollonian circle packing,

See Fig. 4 and 8 for examples of Apollonian circle packings where each circle is labeled by its curvature (that is, the reciprocal of its radius). There are also Apollonian packings which spread all over the plane as well as spread all over to the half plane. As circles in these packings would become enormously large after a few first generations, it is harder to draw them on paper.

There are many natural questions about Apollonian circle packings either from the number theoretic or the geometric point of view and we refer to the series of papers by Graham, Lagarias, Mallows, Wilks, and Yan especially [30] [29], and [17] as well as the letter of Sarnak to Lagarias [64] which inspired the author to work on the topic personally. Also see a more recent article [62].

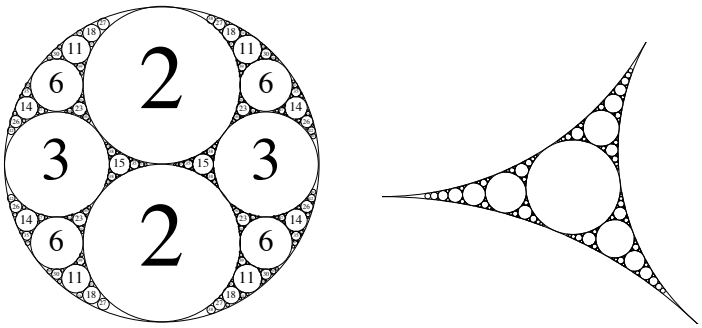


Figure 4. A bounded Apollonian circle packing and the Apollonian packing of a triangular region

To find the symmetry group of a given Apollonian packing $\mathcal{P}$, we consider the dual circles to any fixed four mutually tangent circles (see Fig. 3 where the dotted circles are the dual circles to the solid circles). Inversion with respect to each dual circle fixes three circles that the dual circle crosses perpendicularly and interchanges two circles tangent to those three circles. Hence the group, say, $\Gamma(\mathcal{P})$, generated by the four inversions with respect to the dual circles preserves the packing $\mathcal{P}$ and there are four $\Gamma(\mathcal{P})$ orbits of circles in $\mathcal{P}$.

As the fundamental domain of $\Gamma(\mathcal{P})$ in $\mathbb{H}^3$ can be taken to be the exterior of the four hemispheres above the dual circles in $\mathbb{H}^3$, $\Gamma(\mathcal{P})$ is geometrically finite. It is known that the limit set of $\Gamma(\mathcal{P})$ coincides precisely with the residual set of $\mathcal{P}$ and hence the critical exponent of $\Gamma(\mathcal{P})$ is equal to the Hausdorff dimension of the residual set of $\mathcal{P}$, which is approximately

$$\alpha = 1.30568(8)$$

due to C. McMullen [48] (note that as any two Apollonian packings are equivalent to each other by a Mobius transformation, α is independent of $\mathcal{P}$). In particular it follows that $\Gamma(\mathcal{P})$ is Zariski dense in the real algebraic group $\mathrm{PSL}_2(\mathbb{C})$ and hence we deduce the following from Theorem 3.3 and the remark following it:

Corollary 3.5 ([51]). *Let $\mathcal{P}$ be an Apollonian circle packing. For any bounded region E of $\mathbb{C}$ whose boundary is a countable union of real algebraic curves, we have*

$$N_T(\mathcal{P}, E) \sim \frac{\mathrm{sk}_{\Gamma_{\mathcal{P}}}(\mathcal{P})}{\alpha \cdot |m^{\mathrm{BMS}}_{\Gamma_{\mathcal{P}}}|} \cdot \omega_{\Gamma_{\mathcal{P}}}(E) \cdot T^{\alpha} \quad \text{as } T \to \infty$$

where $\Gamma_{\mathcal{P}} := \Gamma(\mathcal{P}) \cap \mathrm{PSL}_2(\mathbb{C})$.

Remark 3.6. 1. In the cases when $\mathcal{P}$ is bounded and E is the largest disk in such $\mathcal{P}$, and when $\mathcal{P}$ lies between two parallel lines and E is the whole

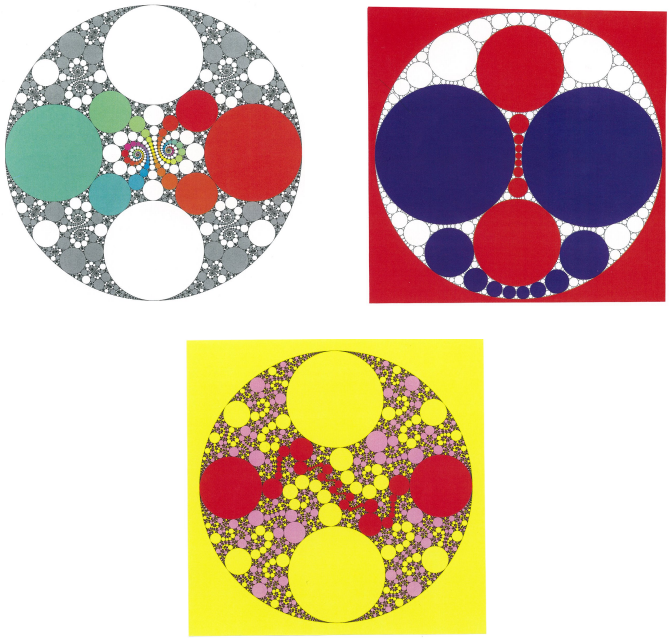


Figure 5. Limit sets of Schottky groups (reproduced with permission from Indra's Pearls, by D.Mumford, C. Series and D. Wright, copyright Cambridge University Press 2002).

period (see Fig. 8), the above asymptotic was previously obtained in [37] with a less explicit description of the main term.

2. Corollary 3.5 applies to any triangular region $\mathcal{T}$ (see Fig. 4) of an Apollonian circle packing.

3.2. More circle packings.

3.2.1. Counting circles in the limit set $\Lambda(\Gamma)$. If $\Gamma\backslash\mathbb{H}^3$ is a hyperbolic 3 manifold with boundary being totally geodesic, then Γ is automatically geometrically finite [34] and $\Omega(\Gamma)$ is a union of countably many disjoint open disks. Hence Theorem 3.3 applies to counting these open disks in $\Omega(\Gamma)$ with respect to the curvature, provided there are infinitely many such. The picture of a Sierpinski curve in Fig. 1 is a special case of this (so are Apollonian circle packings). More precisely, if Γ denotes the group generated by reflections in the sides of a unique regular tetrahedron whose convex core is bounded by four $\frac{\pi}{4}$ triangles and by four right hexagons, then the residual set of a Sierpinski curve in Fig. 1 coincides with $\Lambda(\Gamma)$ (see [47] for details), and it is known to be homeomorphic to the well-known Sierpinski carpet by a theorem of Claytor [9].

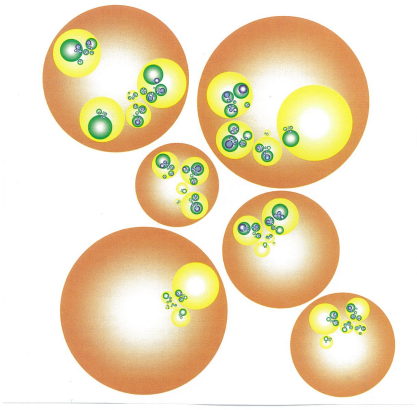


Figure 6. Schottky dance (reproduced with permission from Indra's Pearls, by D.Mumford, C. Series and D. Wright, copyright Cambridge University Press 2002)

Three pictures in Fig. 5 can be found in the beautiful book *Indra's pearls* by Mumford, Series and Wright [50] and the residual sets are the limit sets of some (geometrically finite) Schottky groups and hence our theorem applies to counting circles in those pictures.

3.2.2. Schottky dance. Other kinds of examples are obtained by considering the images of Schottky disks under Schottky groups. Take $k \geq 1$ pairs of mutually disjoint closed disks $\{(D_i, D_i') : 1 \leq i \leq k\}$ in $\mathbb{C}$ and choose Möbius transformations γ_i which maps D_i and D_i' and sends the interior of D_i to the exterior of D_i', respectively. The group, say, Γ, generated by $\{\gamma_i : 1 \leq i \leq k\}$ is called a Schottky group of genus k (cf. [42, Sec. 2.7]). The Γ-orbits of the disks nest down onto the limit set $\Lambda(\Gamma)$ which is totally disconnected. If we denote by $\mathcal{P}$ the union $\cup_{i=1}^{k}(\Gamma(C_i) \cup \Gamma(C_i'))$ where C_i and C_i' are the boundaries of D_i and D_i' respectively, $\mathcal{P}$ is locally finite, as the nesting disks will become smaller and smaller (cf. [50, 4.5]). The common exterior of hemispheres above the initial disks D_i and D_i', $1 \leq i \leq k$, is a fundamental domain for Γ in the upper half-space model $\mathbb{H}^3$, and hence Γ is geometrically finite. Since $\mathcal{P}$ consists of disjoint circles, Theorem 3.3 applies to $\mathcal{P}$. For instance, see Fig. 6 ([50, Fig. 4.11]). One can find many more explicit circle packings in [50] to which Theorem 3.3 applies.

4. Circle Packings on the Sphere

In the unit sphere $\mathbb{S}^2 = \{x^2+y^2+z^2 = 1\}$ with the Riemannian metric induced from $\mathbb{R}^3$, the distance between two points is simply the angle between the rays connecting them to the origin $o = (0, 0, 0)$.

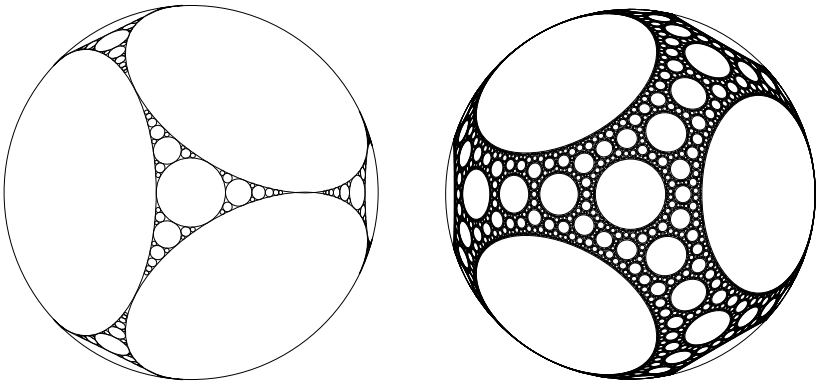


Figure 7. Apollonian packing and Sierpinski curve on the sphere (by C. McMullen)

Let $\mathcal{P}$ be a circle packing on the sphere $\mathbb{S}^2$, i.e., a union of circles. The spherical curvature of a circle C in $\mathbb{S}^2$ is given by

$$\mathrm{Curv}_S(C) = \cot\theta(C)$$

where $0 < \theta(C) \le \pi/2$ is the spherical radius of C, that is, the half of the visual angle of C from the origin o. We suppose that $\mathcal{P}$ is infinite and locally finite in the sense that there are only finitely many circles in $\mathcal{P}$ of spherical curvature at most T for any fixed $T > 0$.

For a region E of $\mathbb{S}^2$, we set

$$N_T(\mathcal{P}, E) := \#\{C \in \mathcal{P} : C \cap E \neq \emptyset,\ \ \mathrm{Curv}_S(C) < T\}.$$

We consider the Poincare ball model $\mathbb{B} = \{x_1^2 + x_2^2 + x_3^2 < 1\}$ of the hyperbolic 3 space with the metric d given by $\frac{2\sqrt{dx_1^2+dx_2^2+dx_3^2}}{1-(x_1^2+x_2^2+x_3^2)}$. Note that the geometric boundary of $\mathbb{B}$ is $\mathbb{S}^2$ and that for any circle C in $\mathbb{S}^2$, we have

$$\sin\theta(C) = \frac{1}{\cosh d(\hat{C}, o)}$$

where $\hat{C} \subset \mathbb{B}$ is the convex hull of C. As both $\sin\theta$ and $\cosh d$ are monotone functions for $0 \le \theta \le \pi/2$ and $d \ge 0$ respectively, understanding $N_T(\mathcal{P}, E)$ is equivalent to investigating the number of Euclidean hemispheres on $\mathbb{B}$ meeting the ball of hyperbolic radius T based at o whose boundaries are in $\mathcal{P}$ and intersect E.

Let G denote the orientation preserving isometry group of $\mathbb{B}$.

Theorem 4.1 ([52]). *Let Γ be a non-elementary geometrically finite discrete subgroup of G and $\mathcal{P} = \cup_{i\in I}\Gamma(C_i)$ be an infinite, locally finite, and Γ-invariant circle packing on the sphere $\mathbb{S}^2$ with finitely many Γ-orbits.*

Suppose one of the following conditions hold:

1. Γ *is convex co-compact;*
2. *all circles in* $\mathcal{P}$ *are mutually disjoint;*
3. $\cup_{i\in I} C_i^\circ \subset \Omega(\Gamma)$ *where* C_i° *denotes the interior of* C_i.

Then for any Borel subset $E \subset \mathbb{S}^2$ *whose boundary is of zero Patterson-Sullivan measure,*

$$N_T(\mathcal{P}, E) \sim \frac{\mathrm{sk}_\Gamma(\mathcal{P}) \cdot \nu_o(E)}{\delta_\Gamma \cdot |m_\Gamma^{\mathrm{BMS}}|} \cdot 2^{\delta_\Gamma} \cdot T^{\delta_\Gamma} \quad \text{as } T \to \infty$$

where $0 < \mathrm{sk}_\Gamma(\mathcal{P}) < \infty$ *is defined in Def. 3.2.*

5. Integral Apollonian Packings: Primes and Twin Primes

A circle packing $\mathcal{P}$ is called *integral* if the curvatures of all circles in $\mathcal{P}$ are integral. One of the special features of Apollonian circle packings is the abundant existence of *integral Apollonian circle packings*.

Descartes noted in 1643 (see [10]) that a quadruple (a, b, c, d) of real numbers can be realized as curvatures of four mutually tangent circles in the plane (oriented so that their interiors are disjoint) if and only if it satisfies

$$2(a^2 + b^2 + c^2 + d^2) - (a + b + c + d)^2 = 0. \tag{5.1}$$

Usually referred to as the Descartes circle theorem, this theorem implies that if the initial four circles in an Apollonian circle packing $\mathcal{P}$ in the plane have integral curvatures, then $\mathcal{P}$ is an integral packing, as observed by Soddy in 1937 [70]. The Descartes circle theorem provides an integral Apollonian packing for every integral solution of the quadratic equation (5.1) and indeed there are infinitely many distinct integral Apollonian circle packings.

Let $\mathcal{P}$ be an integral Apollonian circle packing. We can deduce from the existence of the lower bound for the non-zero curvatures in $\mathcal{P}$ that such $\mathcal{P}$ is either bounded or lies between two parallel lines. We assume that $\mathcal{P}$ is primitive, that is, the greatest common divisor of curvatures is one.

Calling a circle with a prime curvature a prime and a pair of tangent prime circles a twin prime, Sarnak showed:

Theorem 5.2 ([64]). *There are infinitely many primes, as well as twin primes, in* $\mathcal{P}$.

For $\mathcal{P}$ bounded, denote by $\pi^{\mathcal{P}}(T)$ the number of prime circles in $\mathcal{P}$ of curvature at most T, and by $\pi_2^{\mathcal{P}}(T)$ the number of twin prime circles in $\mathcal{P}$ of curvatures at most T. For $\mathcal{P}$ congruent to the packing in Fig. 8, we alter the

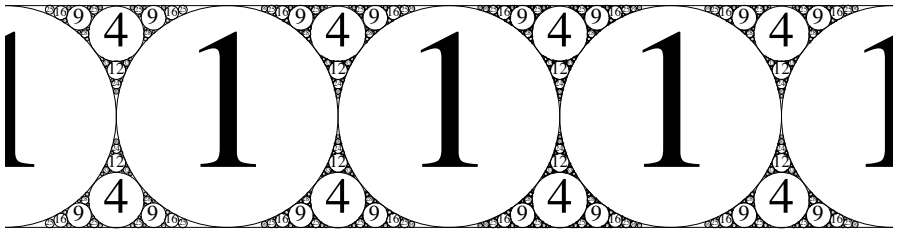


Figure 8. An Apollonian circle packing between two parallel lines.

definition of $\pi^{\mathcal{P}}(T)$ and $\pi_2^{\mathcal{P}}(T)$ to count prime circles in a fixed period. Sarnak showed [64] that

$$\pi^{\mathcal{P}}(T) \gg \frac{T}{(\log T)^{3/2}}.$$

Recently Bourgain, Gamburd and Sarnak ([3] and [4]) obtained a uniform spectral gap for the family of congruence subgroups $\Gamma(q) = \{\gamma \in \Gamma : \gamma \equiv 1 \pmod{q}\}$, q square-free, of any finitely generated subgroup Γ of $\mathrm{SL}_2(\mathbb{Z})$ provided $\delta_\Gamma > 1/2$. This theorem extends to a Zariski dense subgroup Γ of $\mathrm{SL}_2(\mathbb{Z}[i])$ and its congruence subgroups over square free ideals of $\mathbb{Z}[i]$ if $\delta_\Gamma > 1$.

Denoting by Q the Descartes quadratic form

$$Q(a,b,c,d) = 2(a^2+b^2+c^2+d^2) - (a+b+c+d)^2,$$

the approach in [37] for counting circles in Apollonian circle packings which are either bounded or between two parallel lines is based on the interpretation of such circle counting problem into the counting problem for $w\Gamma \cap B_T^{\max}$ where $\Gamma < O_Q(\mathbb{Z})$ is the so-called Apollonian group, $w \in \mathbb{Z}^4$ with $Q(w) = 0$ and $B_T^{\max}$ denotes the maximum norm ball in $\mathbb{R}^4$.

Using the spin double cover $\mathrm{Spin}_Q \to \mathrm{SO}_Q$ and the isomorphism $\mathrm{Spin}_Q(\mathbb{R}) = \mathrm{SL}_2(\mathbb{C})$, we use the aforementioned result of Bourgain, Gamburd and Sarnak to obtain a smoothed counting for $w\Gamma(q) \cap B_T$ with a uniform error term for the family of square-free congruence subgroups $\Gamma(q)$'s where B_T is the Euclidean norm ball. This is a crucial ingredient for the Selberg's upper bound sieve, which is used to prove the following:

Theorem 5.3 ([37]). *As $T \to \infty$,*

$$\pi^{\mathcal{P}}(T) \ll \frac{T^\alpha}{\log T}, \quad \textit{and} \quad \pi_2^{\mathcal{P}}(T) \ll \frac{T^\alpha}{(\log T)^2}$$

where $\alpha = 1.30568(8)$ is the residual dimension of $\mathcal{P}$.

Remark 5.4. 1. Modulo 16, the Descartes equation (5.1) has no solutions unless two of the curvatures are even and the other two odd. In particular, there are no "triplet primes" of three mutually tangent circles, all having odd prime curvatures.

2. We can also use the methods in [37] to give lower bounds for almost primes in a packing. A circle in $\mathcal{P}$ is called *R-almost prime* if its curvature is the product of at most R primes. Similarly, a pair of tangent circles is called *R-almost twin prime* if both circles are R-almost prime. Employing Brun's combinatorial sieve, our methods show the existence of $R_1, R_2 > 0$ (unspecified) such that the number of R_1-almost prime circles in $\mathcal{P}$ whose curvature is at most T is $\asymp \frac{T^\alpha}{\log T}$,[1] and that the number of pairs of R_2-almost twin prime circles whose curvatures are at most T is $\asymp \frac{T^\alpha}{(\log T)^2}$.

3. A suitably modified version of Conjecture 1.4 in [3], a generalization of Schinzel's hypothesis, implies that for some $c, c_2 > 0$,
$$\pi^{\mathcal{P}}(T) \sim c \cdot \frac{T^\alpha}{\log T} \quad \text{and} \quad \pi_2^{\mathcal{P}}(T) \sim c_2 \cdot \frac{T^\alpha}{(\log T)^2}.$$
The constants c and c_2 are detailed in [24].

4. Recently Bourgain and Fuchs [2] showed that in a given bounded integral Apollonian packing $\mathcal{P}$, the growth of the number of *distinct* curvatures at most T is at least $c \cdot T$ for some $c > 0$.

5. The spherical Soddy-Gossett theorem says (see [38]) that the quadruple (a, b, c, d) of spherical curvatures of four mutually tangent circles in $\mathcal{P}$ satisfies
$$2(a^2 + b^2 + c^2 + d^2) - (a + b + c + d)^2 = -4.$$
This theorem implies again that there are infinitely many *integral* spherical Apollonian circle packings, that is, the spherical curvature of every circle is integral. It will be interesting to have results analogous to Theorems 5.2 and 5.3 for integral spherical Apollonian packings.

6. Equidistribution in Geometrically Finite Hyperbolic Manifolds

Let G be the identity component of the group of isometries of $\mathbb{H}^n$ and $\Gamma < G$ be a non-elementary geometrically finite discrete subgroup.

We have discussed that the Bowen-Margulis-Sullivan measure is a finite measure on the unit tangent bundle $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n)$ which is mixing for the geodesic flow. Another measure playing an important role in studying the dynamics of flows on $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n)$ is the following Burger-Roblin measure.

Burger-Roblin measure: The Burger-Roblin measure m_Γ^{BR} is the induced measure on $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n)$ of the following Γ-invariant measure on $\mathrm{T}^1(\mathbb{H}^n)$:
$$d\tilde{m}^{\mathrm{BR}}(u) = e^{(n-1)\beta_{u^+}(x,\pi(u))}\, e^{\delta_\Gamma \beta_{u^-}(x,\pi(u))}\, dm_x(u^+) d\nu_x(u^-) dt$$

[1] By $f(T) \asymp g(T)$, we mean $g(T) \ll f(T) \ll g(T)$.

where m_x denotes the probability measure on the boundary $\partial_\infty(\mathbb{H}^n)$ invariant under the maximal compact subgroup $\mathrm{Stab}_G(x_0)$. For any x and $x_0 \in \mathbb{H}^n$, we have $dm_x(\xi) := e^{-(n-1)\beta_\xi(x,x_0)} dm_{x_0}(\xi)$ and it follows that this definition of m_Γ^{BR} is independent of the choice of $x \in \mathbb{H}^n$.

Burger [8] showed that for a convex cocompact hyperbolic surface with δ_Γ at least $1/2$, this is a unique ergodic horocycle invariant measure up to homothety. Roblin [59] extended Burger's result in much greater generality, for instance, including all non-elementary geometrically finite hyperbolic manifolds.

The name of the Burger-Roblin measure was first suggested by Shah and the author in [37] and [53] in recognition of this important classification result.

We note that the total mass $|m_\Gamma^{\mathrm{BR}}|$ is finite only when $\delta_\Gamma = n-1$ (or equivalently only when Γ is a lattice in G) and is supported on the set $\{u \in \mathrm{T}^1(\Gamma\backslash\mathbb{H}^n) : u^- \in \Lambda(\Gamma)\}$.

Let $S^\dagger \subset \mathrm{T}^1(\mathbb{H}^n)$ be one of the following:

1. an unstable horosphere;
2. the oriented unit normal bundle of a codimension one totally geodesic subspace of $\mathbb{H}^n$
3. the set of outward normal vectors to a (hyperbolic) sphere in $\mathbb{H}^n$.

We consider the following measures on $\mathrm{Stab}_\Gamma(S^\dagger)\backslash S^\dagger$:

$$d\mu_{S^\dagger}^{\mathrm{Leb}}(s) = e^{(n-1)\beta_{s^+}(x,\pi(s))} dm_x(s^+), \quad d\mu_{S^\dagger}^{\mathrm{PS}}(s) = e^{\delta_\Gamma \beta_{s^+}(x,\pi(s))} d\nu_x(s^+)$$

for any $x \in \mathbb{H}^n$.

Denote by p the canonical projection $\mathrm{T}^1(\mathbb{H}^n) \to \mathrm{T}^1(\Gamma\backslash\mathbb{H}^n) = \Gamma\backslash\,\mathrm{T}^1(\mathbb{H}^n)$.

Theorem 6.1 ([53]). *For $\psi \in C_c(\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n))$ and any relatively compact subset $\mathcal{O} \subset p(S^\dagger)$ with $\mu_{S^\dagger}^{\mathrm{PS}}(\partial(\mathcal{O})) = 0$,*

$$e^{(n-1-\delta_\Gamma)t} \cdot \int_{\mathcal{O}} \psi(g^t(s))\, d\mu_{S^\dagger}^{\mathrm{Leb}}(s) \sim \frac{\mu_{S^\dagger}^{\mathrm{PS}}(\mathcal{O}_*)}{\delta_\Gamma \cdot |m_\Gamma^{\mathrm{BMS}}|} \cdot m_\Gamma^{\mathrm{BR}}(\psi) \quad \text{as } t \to \infty$$

where

$$\mathcal{O}_* = \{s \in \mathcal{O} : s^+ \in \Lambda(\Gamma)\}.$$

Definition 6.2. For a hyperbolic subspace $S = \mathbb{H}^{n-1} \subset \mathbb{H}^n$, we say that a parabolic fixed point $\xi \in \Lambda(\Gamma) \cap \partial_\infty(\mathbb{H}^{n-1})$ of Γ is *internal* if any parabolic element $\gamma \in \Gamma$ fixing ξ preserves $\mathbb{H}^{n-1}$.

Recalling the notation π for the canonical projection from $\mathrm{T}^1(\mathbb{H}^n)$ to $\mathbb{H}^n$, we set $S = \pi(S^\dagger)$.

Theorem 6.3 ([53]). *We assume that the projection map* $\mathrm{Stab}_\Gamma(S)\backslash S \to \Gamma\backslash\mathbb{H}^n$ *is proper. In the case when* S *is a codimension one totally geodesic subspace, we also assume that every parabolic fixed point of* Γ *in the boundary of* S *is internal.*

For $\psi \in C_c(\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n))$,

$$e^{(n-1-\delta_\Gamma)t}\cdot\int_{p(S^\dagger)}\psi(g^t(s))\,d\mu^{\mathrm{Leb}}_{S^\dagger}(s)\sim\frac{\mu^{\mathrm{PS}}_{S^\dagger}(S^\dagger_*)}{\delta_\Gamma\cdot|m^{\mathrm{BMS}}_\Gamma|}\cdot m^{\mathrm{BR}}_\Gamma(\psi)\quad \text{as } t\to\infty$$

where

$$S^\dagger_* = \{s\in p(S^\dagger) : s^+\in\Lambda(\Gamma)\}.$$

We have $0\le\mu^{\mathrm{PS}}_{S^\dagger}(S^\dagger_*)<\infty$, *and* $\mu^{\mathrm{PS}}_{S^\dagger}(S^\dagger_*)=0$ *may happen only when* S *is totally geodesic.*

It can be shown by combining results of [11] and [45] that in a finite volume space $\Gamma\backslash\mathbb{H}^n$, the properness of the projection map $\mathrm{Stab}_\Gamma(S)\backslash S\to\Gamma\backslash\mathbb{H}^3$ implies that $\mathrm{Stab}_\Gamma(S)\backslash S$ is of finite volume as well, except for the case when $n=2$ and S is a proper geodesic in $\mathbb{H}^2$ connecting two parabolic fixed points of a lattice $\Gamma<\mathrm{PSL}_2(\mathbb{R})$.

When both $\Gamma\backslash\mathbb{H}^n$ and $\mathrm{Stab}_\Gamma(S)\backslash S$ are of finite volume, we have $n-1=\delta_\Gamma$ and both m^{BMS}_Γ and m^{BR}_Γ are finite invariant measures and $\mu^{\mathrm{PS}}_{S^\dagger}=\mu^{\mathrm{Leb}}_{S^\dagger}$ (up to a constant multiple). In this case, Theorem 6.3 is due to Sarnak [63] for the closed horocycles for $n=2$. The general case is due to Duke, Rudnick and Sarnak [14] and Eskin and McMullen [19] gave a simpler proof of Theorem 6.3, based on the mixing property of the geodesic flow of a finite volume hyperbolic manifold. The latter proof, combined with a strengthened version of the wavefront lemma [28], also works for proving Theorem 6.1. We remark that the idea of using mixing in this type of problem goes back to the 1970 thesis of Margulis [46] (see also [31, Appendix]). Eskin, Mozes and Shah [20] and Shah [69] provided yet another different proofs using the theory of unipotent flows. When both $\Gamma\backslash\mathbb{H}^n$ and $\mathrm{Stab}_\Gamma(S)\backslash S$ are of finite volume, Theorem 6.1 easily implies Theorem 6.3 but not conversely.

In the case when $S^\dagger$ is a horosphere, Theorem 6.1 was obtained in [59], and Theorem 6.3 was proved in [37] when $\delta_\Gamma>(n-1)/2$ with a different interpretation of the main term.

Remark 6.4.

1. The condition on the internality of all parabolic fixed points of Γ in the boundary of S is crucial, as $\mu^{\mathrm{PS}}_{S^\dagger}(S^\dagger_*)=\infty$ otherwise. This can already be seen in the level of a lattice: take $\Gamma=\mathrm{SL}_2(\mathbb{Z})$ and let S be the geodesic connecting 0 and ∞ in the upper half space model. Then any upper triangular matrix in Γ fixes ∞ but does not stabilize S. Indeed the length of the image of S in $\Gamma\backslash\mathbb{H}^2$ is infinite.

2. In proving Theorem 3.3, we count circles in $\mathbb{C}$ by counting the corresponding Euclidean hemispheres in $\mathbb{H}^3$. As the Euclidean hemispheres

are totally geodesic hyperbolic planes, this amounts to understanding the distribution of a Γ-orbit of a totally geodesic hyperbolic plane in $\mathbb{H}^3$. The equidistribution theorem we use here is Theorem 6.3 for S a hyperbolic plane.

3. More classical applications of the equidistribution theorem such as Theorem 6.3 can be found in the point counting problems of Γ-orbits in various spaces.

   For a Γ-orbit in the hyperbolic space $\mathbb{H}^n$, the orbital counting in Riemannian balls was obtained Lax-Phillips [39] for $\delta_\Gamma > \frac{n-1}{2}$ and by Roblin [59] in general.

   Extending the work of Duke, Rudnick and Sarnak [14] and of Eskin and McMullen [19] for Γ lattices, we obtain in [53], for any geometrically finite group Γ of G, the asymptotic of the number of vectors of norm at most T lying in a discrete orbit $w\Gamma$ of a quadric

   $$F(x_1, \cdots, x_{n+1}) = y$$

   for a real quadratic form F of signature $(n, 1)$ and any $y \in \mathbb{R}$ (when $y > 0$, there is an extra assumption on w not being Γ strongly parabolic. See [53] for details). When $y = 0$ and $n = 2, 3$, special cases of this result were obtained in [35], [37] and [36] under the condition $\delta_\Gamma > (n-1)/2$. Based on the Descartes circle theorem, this result in [37] was used to prove Theorem 3.5 for the bounded Apollonian packings. In [40], a Γ-orbit in the geometric boundary is shown to be equidistributed with respect to the Patterson-Sullivan measure, extending the work [25] for the lattice case.

4. For $\psi \in C_c(\Gamma\backslash\mathbb{H}^n)$, we have

   $$m_\Gamma^{\mathrm{BR}}(\psi) = \langle \psi, \phi_\Gamma \rangle := \int_{\Gamma\backslash\mathbb{H}^n} \psi(x) \cdot \phi_\Gamma(x)\, dm^{\mathrm{Leb}}(x)$$

   where $\phi_\Gamma(x) = |\nu_x|$ is the positive eigenfunction of the Laplace operator on $\Gamma\backslash\mathbb{H}^n$ with eigenvalue $\delta_\Gamma(n - 1 - \delta_\Gamma)$ and

   $$dm^{\mathrm{Leb}}(u) = e^{(n-1)\beta_{u^+}(x,\pi(u))}\, e^{(n-1)\beta_{u^-}(x,\pi(u))}\, dm_x(u^+) dm_x(u^-) dt$$

   for any $x \in \mathbb{H}^n$. Hence Theorem 6.3 says that for $\psi \in C_c(\Gamma\backslash\mathbb{H}^n)$,

   $$e^{(n-1-\delta_\Gamma)t} \cdot \int_{p(S^\dagger)} \psi(\pi(g^t(s)))\, d\mu_{S^\dagger}^{\mathrm{Leb}}(s) \sim \frac{\mu_{S^\dagger}^{\mathrm{PS}}(S_*^\dagger)}{\delta_\Gamma \cdot |m_\Gamma^{\mathrm{BMS}}|} \cdot \langle \psi, \phi_\Gamma \rangle \quad \text{as } t \to \infty. \tag{6.5}$$

   When $\delta_\Gamma > (n-1)/2$, $\phi_\Gamma \in L^2(\Gamma\backslash\mathbb{H}^n)$ and its eigenvalue $\delta_\Gamma(n - 1 - \delta_\Gamma)$ is isolated in the L^2-spectrum of the Laplace operator [39]. It will be

desirable to obtain a rate of convergence in (6.5) in terms of the spectral gap of Γ in such cases. For Γ lattices, it was achieved in [14] for $p(S)$ compact and in [1] in general. This was done in the case of a horosphere in [37], which was the main ingredient in the proof of Theorem 5.3. It may be possible to extend the methods of [37] to obtain an error term in general.

7. Further Remarks and Questions

Let G be the identity component of the group of isometries of $\mathbb{H}^n$ and Γ be a geometrically finite group. We further assume that Γ is Zariski dense in G for discussions in this section. When we identify $\mathbb{H}^n$ with G/K for a maximal compact subgroup K, the unit tangent bundle $\mathrm{T}^1(\mathbb{H}^n)$ can be identified with G/M where M is the centralizer in K of a Cartan subgroup, say, A, whose multiplication on the right corresponds to the geodesic flow. The frame bundle of $\mathbb{H}^n$ can be identified with G and the frame flow on the frame bundle is given by the multiplications by elements of A on the right.

We have stated the equidistribution results in section 6 in the level of the unit tangent bundle $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n)$. As the frame bundle is a homogeneous space of G unlike the unit tangent bundle, it is much more convenient to work in the frame bundle. Fortunately, as observed in [23], the frame flow is mixing on $\Gamma\backslash G$ with respect to the lift from $\Gamma\backslash G/M$ to $\Gamma\backslash G$ of the Bowen-Margulis-Sullivan measure. Using this, we can extend Theorems 6.1 and 6.3 to the level of the frame bundle $\Gamma\backslash G$. It seems that the classification theorem of Burger and Roblin can also be extended: for a horospherical group N, any locally finite N-invariant ergodic measure on $\Gamma\backslash G$ is either supported on a closed N-orbit or the lift of the Burger-Roblin measure (we caution here that a locally finite N-invariant measure supported on a closed N-orbit need not be a finite measure unlike the Γ-lattice cases).

In analogy with Ratner's theorem [56], [57], we propose the following problems: let U be a one-parameter unipotent subgroup, or more generally a subgroup generated by unipotent one parameter subgroups of G:

1. [Measure rigidity] Classify all locally finite Borel U-invariant ergodic measures on $\Gamma\backslash G$.

2. [Topological rigidity] Classify the closures of U-orbits in $\Gamma\backslash G$.

We remark that as $G = \mathrm{SO}(n,1)$ (up to a local isomorphism) in our set-up, the above topological rigidity for Γ lattices was also obtained by Shah ([68], [67]) based on the apporoach of Margulis ([43], [44]) and of Dani and Margulis [13].

Both questions are known for $n = 2$ due to Burger [8] and Roblin [59], as in this case, there is only one unipotent one-parameter subgroup up to

conjugation, which gives the horocycle flow. Shapira used them to prove equidistribution for non-closed horocycles [65].

It may be a good idea to start with a sampling case when $G = \mathrm{SL}_2(\mathbb{C})$, $U = \mathrm{SL}_2(\mathbb{R})$ and $\Gamma < G$ Zariski dense and geometrically finite.

1. Are there any locally finite $\mathrm{SL}_2(\mathbb{R})$-invariant ergodic measure on $\Gamma\backslash \mathrm{SL}_2(\mathbb{C})$ besides the Haar measure (=the $\mathrm{SL}_2(\mathbb{C})$-invariant measures) and the $\mathrm{SL}_2(\mathbb{R})$-invariant measures supported on closed $\mathrm{SL}_2(\mathbb{R})$ orbits?

2. Is every non-closed $\mathrm{SL}_2(\mathbb{R})$-orbit dense in $\Gamma\backslash \mathrm{SL}_2(\mathbb{C})$?

It seems that the answers are *no* for (1) and *yes* for (2).

References

[1] Yves Benoist and Hee Oh. Effective equidistribution of S-integral points on symmetric varieties, *Preprint (arXiv:0706.1621)*, 2007.

[2] Jean Bourgain and Elena Fuchs. A proof of the positive density conjecture for integer Apollonian circle packings. *Preprint(arXive:1001.3894)*, 2010.

[3] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Affine linear sieve, expanders and sum-product, 2008. *To appear in Inventiones.*

[4] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Generalization of Selberg's theorem and Selberg's sieve. *Preprint (arXive:0912.5021)*, 2009.

[5] B. H. Bowditch. Geometrical finiteness for hyperbolic groups. *J. Funct. Anal.*, 113(2):245–317, 1993.

[6] Rufus Bowen. Periodic points and measures for Axiom A diffeomorphisms. *Trans. Amer. Math. Soc.*, 154:377–397, 1971.

[7] David W. Boyd. The residual set dimension of the Apollonian packing. *Mathematika*, 20:170–174, 1973.

[8] Marc Burger. Horocycle flow on geometrically finite surfaces. *Duke Math. J.*, 61(3):779–803, 1990.

[9] Schieffelin Claytor. Topological immersion of Peanian continua in a spherical surface. *Ann. of Math. (2)*, 35(4):809–835, 1934.

[10] H. S. M. Coxeter. The problem of Apollonius. *Amer. Math. Monthly*, 75:5–15, 1968.

[11] S. G. Dani. On invariant measures, minimal sets and a lemma of Margulis. *Invent. Math.*, 51(3):239–260, 1979.

[12] S. G. Dani. Flows on homogeneous spaces and Diophantine approximation. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pages 780–789, Basel, 1995. Birkhäuser.

[13] S. G. Dani and G. A. Margulis. Orbit closures of generic unipotent flows on homogeneous spaces of $\mathrm{SL}(3, \mathbf{R})$. *Math. Ann.*, 286(1-3):101–128, 1990.

[14] W. Duke, Z. Rudnick, and P. Sarnak. Density of integer points on affine homogeneous varieties. *Duke Math. J.*, 71(1):143–179, 1993.

[15] M. Einsiedler and E. Lindenstrauss. Diagonalizable flows on locally homogeneous spaces and number theory. In *International Congress of Mathematicians. Vol. II*, pages 1731–1759. Eur. Math. Soc., Zürich, 2006.

[16] N. D. Elkies and C. T. McMullen. Gaps in $\sqrt{n}$ mod 1 and ergodic theory. *Duke Math. J.*, 123(1):95–139, 2004.

[17] Nicholas Eriksson and Jeffrey C. Lagarias. Apollonian circle packings: number theory. II. Spherical and hyperbolic packings. *Ramanujan J.*, 14(3):437–469, 2007.

[18] Alex Eskin. Counting problems and semisimple groups. In *Proceedings of the International Congress of Mathematicians, Vol. II (Berlin, 1998)*, number Extra Vol. II, pages 539–552 (electronic), 1998.

[19] Alex Eskin and C. T. McMullen. Mixing, counting, and equidistribution in Lie groups. *Duke Math. J.*, 71(1):181–209, 1993.

[20] Alex Eskin, Shahar Mozes, and Nimish Shah. Unipotent flows and counting lattice points on homogeneous varieties. *Ann. of Math. (2)*, 143(2):253–299, 1996.

[21] Alex Eskin and Hee Oh. Ergodic theoretic proof of equidistribution of Hecke points. *Ergodic Theory Dynam. Systems*, 26(1):163–167, 2006.

[22] Alex Eskin and Hee Oh. Representations of integers by an invariant polynomial and unipotent flows. *Duke Math. J.*, 135(3):481–506, 2006.

[23] L. Flaminio and R. J. Spatzier. Geometrically finite groups, Patterson-Sullivan measures and Ratner's rigidity theorem. *Invent. Math.*, 99(3):601–626, 1990.

[24] E. Fuchs and K. Sanden. Some experiments with integral apollonian circle packings. *Preprint*, 2010.

[25] Alex Gorodnik and Hee Oh. Orbits of discrete subgroups on a symmetric space and the Furstenberg boundary. *Duke Math. J.*, 139(3):483–525, 2007.

[26] Alex Gorodnik and Hee Oh. Rational points on homogeneous varieties and equidistribution of adelic periods (with an appendix by Borovoi), *Preprint (arXiv:0803.1996)*, 2008.

[27] Alex Gorodnik, Hee Oh, and Nimish Shah. Integral points on symmetric varieties and Satake compactifications. *Amer. J. Math.*, 131(1):1–57, 2009.

[28] Alex Gorodnik, Hee Oh, and Nimish Shah. Strong wavefront lemma and counting lattice points in sectors. *Israel. J. Math.*, 176: 419–444, 2010.

[29] Ronald L. Graham, Jeffrey C. Lagarias, Colin L. Mallows, Allan R. Wilks, and Catherine H. Yan. Apollonian circle packings: number theory. *J. Number Theory*, 100(1):1–45, 2003.

[30] Ronald L. Graham, Jeffrey C. Lagarias, Colin L. Mallows, Allan R. Wilks, and Catherine H. Yan. Apollonian circle packings: geometry and group theory. I. The Apollonian group. *Discrete Comput. Geom.*, 34(4):547–585, 2005.

[31] D. Kleinbock and G. A. Margulis. Bounded orbits of nonquasiunipotent flows on homogeneous spaces. In *Sinaĭ's Moscow Seminar on Dynamical Systems*, volume 171 of *Amer. Math. Soc. Transl. Ser. 2*, pages 141–172. Amer. Math. Soc., Providence, RI, 1996.

[32] D. Kleinbock and G. A. Margulis. Flows on homogeneous spaces and Diophantine approximation on manifolds. *Ann. of Math. (2)*, 148(1):339–360, 1998.

[33] B. Klingler and A. Yafaev. On the Andr'e-Oort conjecture. *Preprint*.

[34] Sadayoshi Kojima. Polyhedral decomposition of hyperbolic 3-manifolds with totally geodesic boundary. In *Aspects of low-dimensional manifolds*, volume 20 of *Adv. Stud. Pure Math.*, pages 93–112. Kinokuniya, Tokyo, 1992.

[35] Alex Kontorovich. The hyperbolic lattice point count in infinite volume with applications to sieves. *Duke Math. J.*, 149(1):1–36, 2009.

[36] Alex Kontorovich and Hee Oh. Almost prime Pythagorean triples in thin orbits. *Preprint (arXive:1001.0370)*, 2010.

[37] Alex Kontorovich and Hee Oh. Apollonian circle packings and closed horospheres on hyperbolic 3-manifolds. *Preprint (arXive:0811.2236)*, 2009.

[38] Jeffrey C. Lagarias, Colin L. Mallows, and Allan R. Wilks. Beyond the Descartes circle theorem. *Amer. Math. Monthly*, 109(4):338–361, 2002.

[39] Peter D. Lax and Ralph S. Phillips. The asymptotic distribution of lattice points in Euclidean and non-Euclidean spaces. *J. Funct. Anal.*, 46(3):280–350, 1982.

[40] Seon-Hee Lim and Hee Oh. On the distribution of orbits of geometrically finite hyperbolic groups on the boundary. *Preprint*, 2010.

[41] Alexander Lubotzky and Robert J. Zimmer. Arithmetic structure of fundamental groups and actions of semisimple Lie groups. *Topology*, 40(4):851–869, 2001.

[42] A. Marden. *Outer circles*. Cambridge University Press, Cambridge, 2007. An introduction to hyperbolic 3-manifolds.

[43] Gregory Margulis. Indefinite quadratic forms and unipotent flows on homogeneous spaces. In *Dynamical systems and ergodic theory (Warsaw, 1986)*, volume 23 of *Banach Center Publ.*, pages 399–409. PWN, Warsaw, 1989.

[44] Gregory Margulis. Orbits of group actions and values of quadratic forms at integral points. In *Festschrift in honor of I. I. Piatetski-Shapiro on the occasion of his sixtieth birthday, Part II (Ramat Aviv, 1989)*, volume 3 of *Israel Math. Conf. Proc.*, pages 127–150. Weizmann, Jerusalem, 1990.

[45] Gregory Margulis. Dynamical and ergodic properties of subgroup actions on homogeneous spaces with applications to number theory. In *Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990)*, pages 193–215, Tokyo, 1991. Math. Soc. Japan.

[46] Gregory Margulis. *On some aspects of the theory of Anosov systems*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2004. With a survey by Richard Sharp: Periodic orbits of hyperbolic flows, Translated from the Russian by Valentina Vladimirovna Szulikowska.

[47] C. T. McMullen. Riemann surfaces, dynamics and geometry. *Course notes for Math 275: available at www.math.harvard.edu/ ctm*.

[48] C. T. McMullen. Hausdorff dimension and conformal dynamics. III. Computation of dimension. *Amer. J. Math.*, 120(4):691–721, 1998.

[49] Philippe Michel and Akshay Venkatesh. Equidistribution, L-functions and ergodic theory: on some problems of Yu. Linnik. In *International Congress of Mathematicians. Vol. II*, pages 421–457. Eur. Math. Soc., Zürich, 2006.

[50] David Mumford, Caroline Series, and David Wright. *Indra's pearls*. Cambridge University Press, New York, 2002. The vision of Felix Klein.

[51] Hee Oh and Nimish Shah. The asymptotic distribution of circles in the orbits of Kleinian groups. *Preprint*, 2010.

[52] Hee Oh and Nimish Shah. Counting visible circles on the sphere and Kleinian groups. *Preprint*, 2010.

[53] Hee Oh and Nimish Shah. Equidistribution and counting for orbits of geometrically finite hyperbolic groups. *Preprint (arXive:1001.2096)*, 2010.

[54] John R. Parker. Kleinian circle packings. *Topology*, 34(3):489–496, 1995.

[55] S.J. Patterson. The limit set of a Fuchsian group. *Acta Mathematica*, 136:241–273, 1976.

[56] Marina Ratner. On Raghunathan's measure conjecture. *Ann. of Math. (2)*, 134(3):545–607, 1991.

[57] Marina Ratner. Raghunathan's topological conjecture and distributions of unipotent flows. *Duke Math. J.*, 63(1):235–280, 1991.

[58] Marina Ratner. Interactions between ergodic theory, Lie groups, and number theory. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pages 157–182, Basel, 1995. Birkhäuser.

[59] Thomas Roblin. Ergodicité et équidistribution en courbure négative. *Mém. Soc. Math. Fr. (N.S.)*, (95):vi+96, 2003.

[60] Daniel J. Rudolph. Ergodic behaviour of Sullivan's geometric measure on a geometrically finite hyperbolic manifold. *Ergodic Theory Dynam. Systems*, 2(3-4):491–512 (1983), 1982.

[61] Omri Sarig. Unique ergodicty for infinite measures. To appear in Proc. ICM (2010).

[62] Peter Sarnak. Integral Apollonian packings. *MAA Lecture, 2009, available at www.math.princeton.edu/ sarnak*.

[63] Peter Sarnak. Asymptotic behavior of periodic orbits of the horocycle flow and eisenstein series. *Comm. Pure Appl. Math.*, 34(6):719–739, 1981.

[64] Peter Sarnak. Letter to J. Lagarias, 2007. available at www.math.princeton.edu/ sarnak.

[65] Barbara Schapira. Equidistribution of the horocycles of a geometrically finite surface. *Int. Math. Res. Not.*, (40):2447–2471, 2005.

[66] Nimish Shah. Equidistribution of translated submanifolds on homogeneous spaces and Dirichler's approximation theorem To appear in Proc. ICM (2010).

[67] Nimish Shah. Closures of totally geodesic immersions in manifolds of constant negative curvature. In *Group theory from a geometrical viewpoint (Trieste, 1990)*, pages 718–732. World Sci. Publ., River Edge, NJ, 1991.

[68] Nimish Shah. Uniformly distributed orbits of certain flows on homogeneous spaces. *Math. Ann.*, 289(2):315–334, 1991.

[69] Nimish Shah. Limit distributions of expanding translates of certain orbits on homogeneous spaces. *Proc. Indian Acad. Sci. Math. Sci.*, 106(2):105–125, 1996.

[70] F. Soddy. The bowl of integers and the hexlet. *Nature*, 139:77–79, 1937.

[71] Dennis Sullivan. The density at infinity of a discrete group of hyperbolic motions. *Inst. Hautes Études Sci. Publ. Math.*, (50):171–202, 1979.

[72] Dennis Sullivan. Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups. *Acta Math.*, 153(3–4):259–277, 1984.

[73] Dennis Sullivan. Related aspects of positivity in Riemannian geometry. *J. Differential Geom.*, 25(3):327–351, 1987.

[74] E. Ullmo. Théorie ergodique et géométrie arithmétique. In *Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002)*, pages 197–206, Beijing, 2002. Higher Ed. Press.

[75] E. Ullmo and A. Yafaev. Galois orbits and equidistribution of special subvarieties: towards the Andr'e-Oort conjecture, 2006. Preprint.

[76] Vinayak Vatsal. Special values of L-functions modulo p. In *International Congress of Mathematicians. Vol. II*, pages 501–514. Eur. Math. Soc., Zürich, 2006.

[77] Shing Tung Yau. Harmonic functions on complete Riemannian manifolds. *Comm. Pure Appl. Math.*, 28:201–228, 1975.

[78] Akihiko Yukie. Prehomogeneous vector spaces and ergodic theory. I. *Duke Math. J.*, 90(1):123–147, 1997.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Equidistribution of Translates of Curves on Homogeneous Spaces and Dirichlet's Approximation

Nimish A. Shah*

Abstract

Understanding the limiting distributions of translates of measures on submanifolds of homogeneous spaces of Lie groups leads to very interesting number theoretic and geometric applications. We explore this theme in various generalities, and in specific cases. Our main tools are Ratner's theorems on unipotent flows, nondivergence theorems of Dani and Margulis, and dynamics of linear actions of semisimple groups.

Mathematics Subject Classification (2010). Primary 22E40; Secondary 11J83.

Keywords. Equidistribution, homogeneous flow, unipotent flow, Ratner's Theorem, Dirichlet's approximation, hyperbolic manifold, geodesic flow

1. Introduction

Several problems in number of theory and geometry involve more than one groups of symmetries or invariance in a direct or an indirect manner. Understanding the dynamics associated to interactions between these groups equips us with deeper new insights into these problems. The proof of Oppenheim conjecture on values of quadratic forms at integral points due to Margulis[16] via study of unipotent flows provided great impetus to the approach of solving number theoretic problems via homogeneous flows techniques. The work of Ratner [17, 18] on classification of invariant measures and orbit closures for unipotent flows as conjectured by Raghunathan and Dani [3] has created the foundation for this area. Since than significant progress and success have been

*Department of Mathematics, The Ohio State University, Columbus, OH 43210, USA, and The Tata Institute of Fundamental Research, Mumbai 400005, India.
Research supported in part by Swarnajayanti Fellowship.
E-mail: shah@math.ohio-state.edu.

achieved in this field by several authors in terms of deep number theoretic and dynamical theorems and powerful techniques. We will discuss a class of such results which are based on describing the limit distributions of sequences of translates of smooth measures on submanifolds in homogeneous spaces of Lie groups.

2. Counting Integral Points on Varieties and Translates of Closed Orbits of Subgroups

Let V be an affine algebraic subvariety of $\mathbb{R}^n$ defined over $\mathbb{Q}$. Let B be a bounded open convex set in $\mathbb{R}^{n-1}$ with smooth boundary. For $T > 0$, define

$$N(T, V) = \text{Cardinality}(V \cap \mathbb{Z}^n \cap TB).$$

In general, it is a difficult problem to estimate $N(T, V)$ as $T \to \infty$.

In [9] Duke, Rudnick and Sarnak observed that when V is an orbit of an algebraic semisimple $\mathbb{Q}$-group G acting linearly on $\mathbb{R}^n$, due a theorem of Borel and Harish-Chandra, $V \cap \mathbb{Z}^n$ is a union of finitely many orbits of a finite index subgroup, say Γ, of $G(\mathbb{Z})$. And hence, if $p \in V \cap \mathbb{Z}^n \neq \emptyset$, we want to obtain asymptotic estimate of

$$N(T, \Gamma p) = \text{Cardinality}(\Gamma p \cap TB)$$

as a function of T for large $T > 0$. Recognizing the role of symmetry and invariance groups in this problem, they noted that if H denotes the stabilizer of p, then under some natural conditions we might expect the following limit to hold:

$$\lim_{T\to\infty} \frac{N(T, \Gamma p)}{\text{Vol}_{G/H}(\{gH \in G/H : g \in G,\ gp \in TB\})} = 1, \tag{1}$$

were the G-invariant $\text{Vol}_{G/H}$ on G/H is determined by the choices of Haar measures on G and H such that $\text{Vol}(G/\Gamma) = \text{Vol}(H/H \cap \Gamma) = 1$.

In [9], they verified this limit for affine symmetric varieties V by introducing a counting technique, and relating it to the following equidistribution result.

Theorem 2.1 (Duke-Rudnick-Sarnak). *Let G be a non-compact simple Lie group, and H be a symmetric subgroup of G; that is, H is the fixed point set of an involutive automorphism (for example, a Cartan involution) of G. Let Γ be a lattice in G, and suppose that $H \cap \Gamma$ is a lattice in H. Let μ_G denote the G-invariant probability measure on G/Γ, and μ_H denote the H-invariant probability measure on G/Γ supported on $H\Gamma/\Gamma \cong H/H \cap \Gamma$. Then for any sequence $\{g_i\}$ in G which is divergent modulo H, we have*

$$\int_{g_i H\Gamma/\Gamma} f\, d(g_i\mu_H) := \int_{y\in H\Gamma/\Gamma} f(g_i y)\, d\mu_H(y) \xrightarrow{i\to\infty} \int_{G/\Gamma} f\, d\mu_G,$$

for any bounded continuous function f on G/Γ.

In other words, the sequence of translated measures $g_i\mu_H$ converge to μ_G in the space of probability measures on G/Γ with respect to the weak-$*$ topology.

The proof of this result in [9] is based on deep results of harmonic analysis of $L^2(G/\Gamma)$. Later Eskin and McMullen [10] deduced Theorem 2.1 as a geometric or a Lie theoretic consequence the mixing property of the sequence of g_i-actions on G/Γ.

The above counting problem and the equidistribution theorem, in view of Ratner's theorem[17] on unipotent flows, motivated the following more general result of [11].

Theorem 2.2 (Eskin-Mozes-Shah). *Let G and $H \subset G$ be connected real algebraic groups defined groups over $\mathbb{Q}$ and admitting no nontrivial $\mathbb{Q}$-characters. Let $\Gamma \subset G(\mathbb{Q})$ be a lattice in G. Let μ_G and μ_H be invariant probability measures G/Γ and $H\Gamma/\Gamma$, respectively. Suppose that for a sequence $\{g_i\}$ in G, the sequence of translated measures $g_i\mu_H$ converges to a probability measure λ on G/Γ with respect to the weak-* topology. Then there exists a $\mathbb{Q}$-subgroup L of G containing H and $c \in G$ such that*

(i) $\lambda = c\mu_L$, were μ_L is the L-invariant probability measure on $L\Gamma/\Gamma$; and

(ii) there exist sequences $\{\gamma_i\} \subset \Gamma$ and $c_i \to c$ in G such that $g_iH = c_i\gamma_iH$ and $\gamma_iH \subset L\gamma_i$ for all large i.

Thus any limit measure is algebraically defined, and the obstruction for this measure to be G-invariant can be algebraically explained.

To prove this theorem one shows that except for the case when g_i is bounded modulo $Z(H) \cap \Gamma$, there exists a sequence $X_i \in \mathrm{Lie}(H)$ such that $X_i \to 0$ and $(\mathrm{Ad}\, g_i)X_i \to Y \neq 0$ in $\mathrm{Lie}(G)$, and λ is invariant under the action of the one-parameter subgroup $\{\exp(tY) : t \in \mathbb{R}\}$. Since 0 is the only eigenvalue of Y, the measure λ is invariant under a unipotent one-parameter subgroup. Now Ratner's theorem describing such measures become applicable to this question.

In [11], using the counting technique introduced by Duke, Rudnick, and Sarnak, the above result was used for proving (1) under appropriate conditions for a wide class of varieties V, and in particular, when H is a maximal Q-subgroup of G. For example, we show the following:

Let $p(x) \in \mathbb{Z}[x]$ be an irreducible monic polynomial. Then the cardinality of the set of $n \times n$ integral matrices of norm at most T and having $p(x)$ as the characteristic polynomial is asymptotically equivalent to $cT^{n(n-1)/2}$, where $c > 0$ is a constant which can be described in terms of class number, regulator, and discriminant associated to the number field generated by a root of $p(x)$ (cf. [22]).

2.1. Expanding translates of smooth measures on horospherical leaves. The work of Eskin and McMullen [10] also motivated the following result [21]:

Theorem 2.3 (Shah). *Let G be a noncompact simple Lie group, and $g \in G$ be a semisimple element not contained in a compact subgroup of G . Let $U = \{u \in G : g^{-n}ug^n \to e \text{ as } n \to \infty\}$ denote the expanding horospherical subgroup of g. Let L be a Lie group containing G, and Γ a lattice in L such that Gx_0 is dense in L/Γ, where $x_0 = e\Gamma$. Let λ be a probability measure on U which is absolutely continuous with respect to a Haar measure on U. Let $\bar{\lambda}$ be the pushforward of λ on Gx_0 under the map $h \mapsto hx_0$ from G to L/Γ. Then as $n \to \infty$, $a^n\bar{\lambda}$ converges weakly to μ_L, the L-invariant probability measure on L/Γ. In other words, for any bounded continuous function f on L/Γ,*

$$\lim_{n\to\infty} \int_{h\in U} f(g^n h x_0)\, d\lambda(h) = \int_{L/\Gamma} f\, d\mu_L.$$

The above result can be generalized as follows: Let

$$P^- = \{b \in G : \overline{\{g^n b g^{-n} : n \in \mathbb{N}\}} \text{ is compact}\}$$

denote the stable subgroup for g. Let λ be any probability measure on G such that the pushforward of λ on $P^-\backslash G$ is absolutely continuous. Let $\bar{\lambda}$ denote the pushforward of λ on Gx_0. Then $g^n\bar{\lambda}$ converges weakly to μ_L.

As a special case one generalizes Theorem 2.1 as follows: Let H be a symmetric subgroup of G, λ be a probability measure which is absolutely continuous with respect to a Haar measure on H, and $\bar{\lambda}$ denote the pushforward of λ on Hx_0. Then for any sequence $\{g_i\} \subset G$, which diverges modulo H, the sequence $g_i\bar{\lambda}$ converges weakly to μ_L as $i \to \infty$. This result has interesting consequences to equidistribution of dense orbits of lattices on homogeneous spaces [13, 12].

3. Limits of measures on stretching translates of submanifolds

In view of the results and notation of subsection 2.1, we ask the following question: Let M be an immersed submanifold of U with $\dim(M) < \dim(U)$ and λ be a probability measure on M which is absolutely continuous with respect to a smooth measure on M. Let $\bar{\lambda}$ denote the pushforward of λ on Gx_0. Under what condition on the geometric shape of M we have that $g^n\bar{\lambda} \to \mu_L$ as $n \to \infty$?

3.0.1. An algebraic obstruction to the limit of $g^n\bar{\lambda}$ being equal to μ_L. Define

$$P_L^- = \{b \in L : \overline{\{g^n b g^{-n} : n > 0\}} \text{ is compact}\}.$$

Suppose that H is a proper subgroup of L containing g, and $q \in L$ is such that the orbit Hqx_0 is closed and carries a finite H-invariant measure. Suppose that $M \subset U \cap P_L^- Hq$. Then any weak-* limit of probability measures $g^n\bar{\lambda}$ is a direct integral of measures which are supported on closed sets of the form

$bHqx_0$, where $b \in P_L^-$ is such that $\overline{\{g^n b g^{-n} : n < 0\}}$ is compact. Such limiting measures are concentrated on strictly low dimensional submanifolds of L/Γ.

We ask if this is the only condition on the geometric shape of M. In the remaining article we will show that this is indeed the case in certain specific situations, and obtain new number theoretic and geometric consequences.

3.1. Translates of a finite arc under geodesic flow. Let $G = \mathrm{SO}(n,1)$ and $\{a_t\}$ be a connected maximal $\mathbb{R}$-diagonalizable subgroup of G. Let $P^- = \{b \in G : \overline{\{a_t b a_{-t} : t > 0\}} \text{ is compact}\}$ and U be the corresponding expanding horospherical subgroup of G. Here $P^-\backslash G \cong \mathbb{S}^{n-1}$ and $U \cong \mathbb{R}^{n-1}$, and the map $u \mapsto P^- u$ from U to $P^-\backslash G$ correspond to the inverse-stereographic projection, and the right action of G on $P^-\backslash G \cong \mathbb{S}^{n-1}$ is via conformal transformations. If H is a proper closed subgroup of G containing $\{a_t\}$ and some nontrivial unipotent subgroup, then $P^- H$ correspond to a proper subsphere of $\mathbb{S}^{n-1}$. Therefore $U \cap P^- Hg$ is an affine subspace or a subsphere in $U \cong \mathbb{R}^{n-1}$. In [23] we show the following:

Theorem 3.1 (Shah). *Let $\phi : (0,1) \to U$ be an analytic map such that $\phi(0,1)$ is not contained in a proper subsphere or a proper affine subspace. Then for any lattice Γ in G, $x \in G/\Gamma$ and any bounded continuous function f on G/Γ,*

$$\lim_{t\to\infty} \int_0^1 f(a_t \phi(s) x)\, ds = \int_{G/\Gamma} f \, d\mu_G$$

where μ_G is the G-invariant probability measure on G/Γ.

The above result was generalized for smooth maps in [24]. We can obtain its following geometric application:

Let $\mathbb{H}^n$ denote the hyperbolic n-ball. Let $\Gamma \subset \mathrm{SO}(n,1)$ be a torsion free discrete group of isometries of $\mathbb{H}^n$ such that the hyperbolic manifold $M = \mathbb{H}^n/\Gamma$ has finite Riemannian volume. Let $\pi : T^1(\mathbb{H}^n) \to T^1(M)$ denote the natural quotient map of the unit tangent bundles, and let g_t denote the geodesic flow on $T^1(M)$. For $v \in T^1(\mathbb{H}^n)$, let $v^+ \in \partial\mathbb{H}^n$ denote the end of the directed geodesic starting from v.

Theorem 3.2 (Shah). *Let $\phi : [0,1] \to T^1(\mathbb{H}^n)$ be a continuous map such that the map $s \mapsto \phi(s)^+ : (0,1) \to \partial\mathbb{H}^n$ is C^1 and its derivative $d\phi(s)^+/ds$ is Lipschitz and nonzero for almost all s. Suppose that the set $\{s \in (0,1) : \phi(s)^+ \in S\}$ has zero Lebesgue measure for any proper subsphere $S \subset \partial\mathbb{H}^n$ such that S is the boundary of an isometric copy of $\mathbb{H}^k$ $(2 \le k < n)$ in $\mathbb{H}^n$ whose image on M is a closed subset. Then for any bounded continuous function f on $T^1(M)$,*

$$\lim_{t\to\infty} \int_0^1 f(g_t \pi(\phi(s)))\, ds = \int_{T^1(M)} f \, d\tilde{\mu}_M,$$

where $\tilde{\mu}_M$ is the probability measure on $T^1(M)$ corresponding to the natural Riemannian volume form on $T^1(M)$.

When ϕ is analytic, the condition of the theorem holds if the image of ϕ^+ is not contained in a proper subsphere of $\partial\mathbb{H}^n$.

4. Applications to Diophantine approximation

The above study was also prompted by the following result due to Kleinbock and Margulis [14]: Let $n \geq 2$ and $\Omega := \{g\mathbb{Z}^n : g \in \mathrm{SL}(n,\mathbb{R})\} \cong \mathrm{SL}(n,\mathbb{R})/\mathrm{SL}(n,\mathbb{Z})$ denote the space unimodular lattices in $\mathbb{R}^n$. Given $\epsilon > 0$, define $\Omega(\epsilon) = \{\Lambda \in \Omega : \|\boldsymbol{v}\| \geq \epsilon, \forall \boldsymbol{v} \in \Lambda \smallsetminus \{0\}\}$. Then $\Omega(\epsilon)$ is compact, and $\cup_{\epsilon>0}\Omega(\epsilon) = \Omega$.

For $\boldsymbol{t} = (t_1, \ldots, t_{n-1}) \in \mathbb{R}^{n-1}$ and $\boldsymbol{v} = (v_1, \ldots, v_{n-1}) \in \mathbb{R}^{n-1}$, define

$$a(\boldsymbol{t}) = \begin{bmatrix} e^{t_1+\cdots+t_{n-1}} & & & \\ & e^{-t_1} & & \\ & & \ddots & \\ & & & e^{-t_{n-1}} \end{bmatrix}, \quad u(\boldsymbol{v}) = \begin{bmatrix} 1 & v_1 & \ldots & v_{n-1} \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

Theorem 4.1 (Kleinbock-Margulis). *Let $\phi : (0,1) \to \mathbb{R}^{n-1}$ be a non-degenerate C^n-map; that is, for almost all $t \in (0,1)$, the derivatives $\phi^{(i)}(t)$, $1 \leq i \leq n-1$, span $\mathbb{R}^{n-1}$. Then there exist constants $C > 0$ and $\alpha > 0$ such that*

$$\ell(\{s \in (0,1) : a(\boldsymbol{t})u(\phi(s))\mathbb{Z}^n \notin \Omega(\epsilon)\}) \leq C\epsilon^\alpha, \quad \forall \epsilon > 0, \forall \boldsymbol{t} \in \mathbb{R}^{n-1}_+.$$

Kleinbock and Margulis [14] used this result to settle conjectures on metric properties of diophantine approximation on submanifolds of $\mathbb{R}^n$ due to Mahler, Sprindzuk and Baker.

The result raises the following dynamical question: Let ν denote the push-forward of the Lebesgue measure on $(0,1)$ under the map $s \mapsto u(\phi(s))x_0$ on Ω. Let $\boldsymbol{t}_i \in \mathbb{R}^{n-1}_+$ be a sequence such that all coordinates of $\boldsymbol{t}_i$ tend to infinity. Then as $i \to \infty$, does the measure $a(\boldsymbol{t}_i)\nu$ tend to μ, the unique $\mathrm{SL}(n,\mathbb{R})$-invariant probability measure on Ω?

It was observed by Kleinbock and Weiss [15] that an affirmative answer to this question would resolve a problem proposed by Davenport and Schmidt [7] in the late 60's on non-improvability of Dirichlet's simultaneous approximation theorem. To describe the problem, consider the following definition:

Given $\lambda > 0$ we say that $\boldsymbol{\xi} \in \mathbb{R}^k$ is $\mathrm{DT}(\lambda)$ if for all but finitely many $N \in \mathbb{N}$, there exist $0 \neq \boldsymbol{q} = (q_1, \ldots, q_k) \in \mathbb{Z}^k$ and $p \in \mathbb{Z}$ such that

$$|\boldsymbol{q} \cdot \boldsymbol{\xi} + p| \leq \lambda/N^k \quad \text{and} \quad |q_i| \leq N, \forall i. \tag{2}$$

Similarly, we say that $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k) \in \mathbb{R}^k$ is $\mathrm{DT}'(\lambda)$ if for all but finitely many $N \in \mathbb{N}$, there exist $0 \neq q \in \mathbb{Z}$ and $\boldsymbol{p} \in \mathbb{Z}^k$ such that

$$|q\xi_i + \boldsymbol{p}| \leq \lambda/N, \ \forall i, \quad \text{and} \quad |q| \leq N^k.$$

Dirichlet's simultaneous approximation theorem states that every $\boldsymbol{\xi} \in \mathbb{R}^k$ is $\mathrm{DT}(1)$ and $\mathrm{DT}'(1)$. Davenport and Schmidt [6] showed that for any $\lambda < 1$,

almost every $\boldsymbol{\xi} \in \mathbb{R}^k$ is not $\mathrm{DT}(\lambda)$ and not $\mathrm{DT}'(\lambda)$. In [7] they showed that for almost any $\xi \in \mathbb{R}$ the vector (ξ, ξ^2) is not $\mathrm{DT}(1/4)$. The result was generalized by Baker [1] for points on more general curves on $\mathbb{R}^2$, by Dodson, Rynne and Vickers [8] for points on 'low co-dimensional curved submanifolds' of $\mathbb{R}^n$, by Bugeaud [2] for the curve $(\xi, \xi^2, \dots, \xi^n)$, and by Kleinbock and Weiss [15] for all nondegenerate curves on $\mathbb{R}^k$. In each case, it was proved that almost all points of the parametrized submanifold with respect to the parameter measure are not $\mathrm{DT}(\lambda)$ for some very small value of $\lambda > 0$ depending on the submanifold.

In [25] we provide the following answer to the above problem:

Theorem 4.2 (Shah). *Let B be a ball in $\mathbb{R}^d$ for some $d \geq 1$, and $\phi : B \to \mathbb{R}^k$ be an analytic map whose image is not contained in a proper affine subspace of R^k. Then for almost every $b \in B$, the point $\phi(b)$ is neither $\mathrm{DT}(\lambda)$ nor $\mathrm{DT}'(\lambda)$ for any $\lambda < 1$.*

The above statement is a consequence of the following equidistribution result [25]:

Theorem 4.3 (Shah). *Let L be any Lie group and $\rho : G = \mathrm{SL}(n, \mathbb{R}) \to L$ be a continuous homomorphism. Let Γ be a lattice in L. Let B be a bounded open subset in $\mathbb{R}^d$ ($d \geq 1$). Let $\phi : B \to \mathrm{SL}(n, \mathbb{R})$ be an analytic map such that the image of the first row of this map is not contained in a proper subspace of $\mathbb{R}^n$. Put $a_t = a((t, t, \dots, t)) \in \mathrm{SL}(n, \mathbb{R})$ ($t \in \mathbb{R}$). Let $x \in L/\Gamma$ and suppose that $\rho(G)x$ is dense in L/Γ. Then for a bounded continuous function f on L/Γ,*

$$\lim_{t \to \infty} \frac{1}{\mathrm{Vol}(B)} \int_{b \in B} f(\rho(a_t u(\phi(b))x)\, db = \int_{L/\Gamma} f \, d\mu_L, \tag{3}$$

where db denotes the Lebesgue integral on $\mathbb{R}^d$, and μ_L is the L-invariant probability measure on L/Γ.

4.0.1. Expanding translates of shrinking submanifolds. Fix any $b \in B$ and let B_t denote a ball of radius e^{-t} about b. If B is replaced by the shrinking balls B_t in (3) then we still expect the limiting measure to be μ_L. This has been verified in the case of $n = 3$. This type of result would allows us to deduce the above theorem when ϕ to is a non-degenerate C^n curve as in Theorem 4.1.

4.1. Multiplicative Dirichlet-Minkowski approximation.

The following generalization of Dirichlet's theorem is known as Minkowski's theorem on simultaneous approximation of Linear forms: For $n \geq 2$, let $(\phi_{ij}) \in \mathrm{SL}(n, \mathbb{R})$. Let $\alpha_1, \dots, \alpha_n > 0$ be such that $\alpha_1 \cdots \alpha_n = 1$. Then there exist $x_1, \dots, x_n \in \mathbb{Z}$, not all 0s, such that

$$|\phi_{11}x_1 + \cdots + \phi_{1n}x_n| \leq \alpha_1;\ |\phi_{i1}x_1 + \cdots + \phi_{in}x_n| < \alpha_i \quad (i \geq 2). \tag{4}$$

By putting $\phi_{11} = \cdots = \phi_{nn} = 1$ and $\phi_{ij} = 0$ for $i \geq 2$ and $j \neq i$, we get a multiplicative version Dirichlet's theorem. Now we define the corresponding

λ-version: For $k = n-1$, let $\mathcal{N} \subset \mathbb{N}^k$ be an infinite sequence and $0 < \lambda \leq 1$. We say that $(\xi_1, \ldots, \xi_k) \in \mathbb{R}^k$ is MDT(λ) along $\mathcal{N}$ if for all but finitely many $(N_1, \ldots, N_k) \in \mathcal{N}$, there exist $q_1, \ldots, q_k \in \mathbb{Z}$, not all zero, and $p \in \mathbb{Z}$ such that

$$|p + q_1\xi_1 + \cdots + q_k\xi_k| \leq \lambda/(N_1 N_2 \ldots N_k) \text{ and } |q_i| < N_i, \ \forall i. \tag{5}$$

We also define MDT$'(\lambda)$ in a similar way. Minkowski's result implies that all points are MDT(1) and MDT$'(1)$ along any $\mathcal{N}$.

Kleinbock and Weiss [15] proved that if each coordinate projection of $\mathcal{N}$ is a divergent sequence then almost all $\boldsymbol{\xi} \in \mathbb{R}^k$ are neither MDT(λ) nor MDT$'(\lambda)$ along $\mathcal{N}$ for any $\lambda < 1$. They also showed that given a non-degenerate smooth curve in $\mathbb{R}^k$, there exists a very small $\lambda > 0$ so that for almost every $\boldsymbol{\xi}$ on this curve is not MDT(λ) along $\mathcal{N}$.

For analytic curves not contained in proper affine subspaces of $\mathbb{R}^k$ we extend their result for any $\lambda < 1$ in [26] as follows:

Theorem 4.4 (Shah). *Let $\mathcal{N}$ be an infinite subset of $\mathbb{N}^k$. Let B be an open ball in $\mathbb{R}^d$ and $\phi : B \to \mathbb{R}^k$ be an analytic map whose image is not contained in a proper affine subspace. Then for almost all $b \in B$ with respect to the Lebesgue measure on $\mathbb{R}^d$ and $\lambda < 1$ there exist infinitely many $(N_1, \ldots, N_k) \subset \mathcal{N}$ such that both the following sets of inequalities are simultaneously insoluble:*

$$|q_1\phi_1(b) + \cdots + q_k\phi_k(b) + p| \leq \lambda/(N_1 \cdots N_k), \quad |q_i| \leq N_i \ (\forall i), \tag{6}$$

for $p, q_1, \ldots, q_k \in \mathbb{Z}$, not all zeros; and

$$|q\phi_i(b) + p_i| \leq \lambda N_i^{-1} \ (\forall i), \quad |q| \leq N_1 N_2 \ldots N_k, \tag{7}$$

for $p_1, \ldots, p_k, q \in \mathbb{Z}$, not all zeros.

In particular, $\phi(b)$ is neither MDT(λ) nor MDT$'(\lambda)$ along $\mathcal{N}$ for any $\lambda < 1$ and almost all $b \in B$.

It may be noted that, due to a theorem of Minkowski and Hajosh on critical lattices, the analogue of the above theorem on multiplicative non-improvability along $\mathcal{N}$ fails to hold if we take an unbounded sequence $\mathcal{N}$ contained $(\mathbb{R}_+)^k$ such that one of the coordinates of $\mathcal{N}$ converges to an element of $\mathbb{R} \smallsetminus \mathbb{N}$ (see [26]).

The deductions of the above results are based on the following relation between the approximation inequality and matrix action on the space of unimodular lattices in $\mathbb{R}^{k+1}$ (see[4, 14, 15]); that is, the inequalities (5) are equivalent to

$$\begin{bmatrix} N_1 \cdots N_k & & & \\ & N_1^{-1} & & \\ & & \ddots & \\ & & & N_k^{-1} \end{bmatrix} \begin{bmatrix} 1 & \xi_1 & \cdots & \xi_{n-1} \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \begin{bmatrix} p \\ q_1 \\ \vdots \\ q_k \end{bmatrix} \in [-\lambda, \lambda] \times [-1, 1]^k,$$

or in other words $a(\boldsymbol{t})u(\boldsymbol{\xi})x_0 \in L_\lambda$, where $\boldsymbol{t} = (\log N_1, \ldots, \log N_k)$, $x_0 = \mathbb{Z}^n \in \Omega$, and

$$L_\lambda = \{g\mathbb{Z}^n \in \Omega : g \in \mathrm{SL}(n, \mathbb{R}), \ g\mathbb{Z}^n \cap [-\lambda, \lambda] \times [-1, 1]^k \neq \{0\}\}$$

is the complement of a nonempty open subset of Ω if $0 < \lambda < 1$. In view of this relation, the dynamical result needed to prove theorem 4.4 is as follows [26]: Given an unbounded sequence $\{\boldsymbol{t}_i\}$ in $\mathbb{R}^{n-1}_+$, after permuting coordinates and passing to a subsequence, we will assume that its first m coordinate projections are divergent sequences ($1 \leq m \leq n-1$), and its remaining $(n-1-m)$ coordinate projections are convergent sequences. Let

$$Q = \Big\{(g_{i,j}) \in \mathrm{SL}(n,\mathbb{R}) : \text{for } i > m+1, \ \begin{matrix} g_{i,j} = 0 & \text{if } j \neq i \\ g_{i,i} = 1 & \end{matrix} \Big\}. \tag{8}$$

Then as $i \to \infty$, $a(\boldsymbol{t}_i)Q \to a(\boldsymbol{t}_0)Q$ in $\mathrm{SL}(n,\mathbb{R})/Q$ for some $\boldsymbol{t}_0 \in \mathbb{R}^{n-1}$. In particular, if all coordinates of $\boldsymbol{t}_i$ are divergent then $Q = G$ and $\boldsymbol{t}_0 = 0$.

Theorem 4.5 (Shah). *Let B be a bounded open subset of $\mathbb{R}^d$ ($d < n$). Let $\phi : B \to \mathbb{R}^{n-1}$ be an analytic map whose image is not contained in a proper affine subspace. Let L be a Lie group, $\rho : \mathrm{SL}(n,\mathbb{R}) \to L$ be a continuous homomorphism, and Γ be a lattice in L. Let $\{\boldsymbol{t}_i\}$ be a sequence as above. Let $x \in L/\Gamma$. Then for any bounded continuous function f on L/Γ,*

$$\lim_{i\to\infty} \frac{1}{\mathrm{Vol}(B)} \int_B f(\rho(a(\boldsymbol{t}_i)u(\phi(b)))x)\, db = \int_{y\in Hx} f(\rho(a(\boldsymbol{t}_0))y)\, d\mu_H(y),$$

where H is the smallest closed subgroup of L containing $\rho(Q)$ such that Hx is closed and admits an H-invariant probability measure, say μ_H.

5. Unipotent flows, Linearization and Linear dynamics

To prove the above dynamical results one shows that if λ is the normalized parameter measure on the submanifold $\rho(u(\phi(B)))x$ of L/Γ, which is being translated by a sequence $g_i = \rho(a(\boldsymbol{t}_i))$, and if we prove that $g_i\lambda$ converges to a measure μ on L/Γ, then μ turns out to be a direct integral of finite measures which are invariant under actions unipotent subgroups of G. Due to Ratner's measure classification theorem, if μ is not L-invariant, then μ is strictly positive on the image of a proper algebraic subvariety, say $\mathcal{V}$ of L projected to L/Γ. This variety is right invariant under certain subgroup, say N, containing unipotents and such that $N\Gamma$ is closed. At this stage one applies linearization technique [19, 5, 20] in conjunction with functions of (C,α)-growth (as introduced in [14]) to show that for each $a(\boldsymbol{t}_i)$ there exists $\gamma_i \in L$ stabilizing x such that $\rho(a(\boldsymbol{t}_i)u(\phi(B)))\gamma_i$, a lift of the entire translated trajectory, lives in a thin neighbourhood of the subvariety $\mathcal{V}$ in L modulo N . At this stage we invoke the following new observation of linear dynamical nature, to deduce that there exist some fixed $\gamma \in L$ stabilizing x such that $\rho(a(\boldsymbol{t}_i)u(\phi(B))\gamma$ gets arbitrarily close to $\mathcal{V}$ in L modulo N. The linear dynamical observation, which turns out to be one of the most crucial part of the argument, is as follows [23, 24, 25, 26]:

Theorem 5.1 (Shah). *Let $\phi : (0,1) \to \mathbb{R}^{n-1}$ be a C^1-map such that for some interval $B \subset (0,1)$, $\phi(B)$ is not contained in a proper affine subspace of $\mathbb{R}^{n-1}$. Suppose that $\mathrm{SL}(n,\mathbb{R})$ acts linearly on a finite dimensional vector space V. Let a sequence $\{\boldsymbol{t}_i\}$ and the associated subgroup Q be as in (8). Then for any $v \in V$ which is not fixed by Q, and any compact set $C \subset V$,*

$$a_{\boldsymbol{t}_i} u(\phi(B))v \not\subset C \qquad \text{for all large } i. \tag{9}$$

Note that if v is fixed by Q then $a_{\boldsymbol{t}_i} u(\phi(B))v = a_{\boldsymbol{t}_i} v \to a_{\boldsymbol{t}_0} v$ as $i \to \infty$.

Our proof of this result uses the description of finite dimensional representations of $\mathrm{SL}(2,\mathbb{R})$ to understand the intertwined linear dynamics of various copies of $\mathrm{SL}(2,\mathbb{R})$s and $\mathrm{SL}(m,\mathbb{R})$s sitting in $\mathrm{SL}(n,\mathbb{R})$.

In the case when ϕ is a nondegenerate C^n-map, we expect that (9) will hold even if we put B_i in place of B where B_i's are intervals around some $s \in (0,1)$ shrinking at some specific rate depending on $a(\boldsymbol{t}_i)$. For example, in the case when $\boldsymbol{t}_i = (t_i, \dots, t_i)$ (all same coordinates) then we can shrink B_i (around any s except for finitely many $s \in B$) at the rate of e^{-t_i} as $i \to \infty$, and (9) can be expected to hold.

The basic strategy behind the dynamical theorems of the previous section is that in very general situations, using Ratner's theorem and Linearization techniques we can reduce the equidistribution problem to a problem about 'Dynamics of subgroup actions on finite dimensional linear representations'. At that stage n we need to prove the results that are very similar to Theorem 5.1, possibly with B also shrinking at a very specific rate as $i \to \infty$. Proving a suitable linear dynamical result remains to be the the main difficulty in describing the limiting distributions of stretching translates of submanifolds on homogeneous spaces of very general Lie groups.

Acknowledgements. I would like express my gratitude towards my teachers and mentors M.S. Raghunathan, S.G. Dani, Gopal Prasad, G. A. Margulis and M. Ratner. I am very thankful to my friends, collaborators and colleagues for their support and guidance through various stages of my mathematical career. I would like to thank my wife for her strong support in all situations.

References

[1] R. C. Baker. Dirichlet's theorem on Diophantine approximation. *Math. Proc. Cambridge Philos. Soc.*, 83(1):37–59, 1978.

[2] Yann Bugeaud. Approximation by algebraic integers and Hausdorff dimension. *J. London Math. Soc. (2)*, 65(3):547–559, 2002.

[3] S. G. Dani. Invariant measures of horospherical flows on noncompact homogeneous spaces. *Invent. Math.*, 47(2):101–138, 1978.

[4] S. G. Dani. Divergent trajectories of flows on homogeneous spaces and Diophantine approximation. *J. Reine Angew. Math.*, 359:55–89, 1985.

[5] S. G. Dani and G. A. Margulis. Limit distributions of orbits of unipotent flows and values of quadratic forms. In *I. M. Gelfand Seminar*, pages 91–137. Amer. Math. Soc., Providence, RI, 1993.

[6] H. Davenport and W. M. Schmidt. Dirichlet's theorem on diophantine approximation. II. *Acta Arith.*, 16:413–424, 1969/1970.

[7] H. Davenport and Wolfgang M. Schmidt. Dirichlet's theorem on diophantine approximation. In *Symposia Mathematica, Vol. IV (INDAM, Rome, 1968/69)*, pages 113–132. Academic Press, London, 1970.

[8] M. M. Dodson, B. P. Rynne, and J. A. G. Vickers. Dirichlet's theorem and Diophantine approximation on manifolds. *J. Number Theory*, 36(1):85–88, 1990.

[9] W. Duke, Z. Rudnick, and P. Sarnak. Density of integer points on affine homogeneous varieties. *Duke Math. J.*, 71(1):143–179, 1993.

[10] Alex Eskin and Curt McMullen. Mixing, counting, and equidistribution in Lie groups. *Duke Math. J.*, 71(1):181–209, 1993.

[11] Alex Eskin, Shahar Mozes, and Nimish Shah. Unipotent flows and counting lattice points on homogeneous varieties. *Ann. of Math. (2)*, 143(2):253–299, 1996.

[12] Alex Gorodnik and Barak Weiss. Distribution of lattice orbits on homogeneous varieties. *Geom. Funct. Anal.*, 17(1):58–115, 2007.

[13] Alexander Gorodnik and Hee Oh. Orbits of discrete subgroups on a symmetric space and the Furstenberg boundary. *Duke Math. J.*, 139(3):483–525, 2007.

[14] D. Y. Kleinbock and G. A. Margulis. Flows on homogeneous spaces and Diophantine approximation on manifolds. *Ann. of Math. (2)*, 148(1):339–360, 1998.

[15] Dimitry Kleinbock and Barak Weiss. Dirichlet's theorem on diophantine approximation and homogeneous flows. *Journal of Modern Dynamics (JMD)*, 2(1):43–62, 2008.

[16] G. A. Margulis. Discrete subgroups and ergodic theory. In *Number theory, trace formulas and discrete groups (Oslo, 1987)*, pages 377–398. Academic Press, Boston, MA, 1989.

[17] Marina Ratner. On Raghunathan's measure conjecture. *Ann. of Math. (2)*, 134(3):545–607, 1991.

[18] Marina Ratner. Raghunathan's topological conjecture and distributions of unipotent flows. *Duke Math. J.*, 63(1):235–280, 1991.

[19] Nimish A. Shah. Uniformly distributed orbits of certain flows on homogeneous spaces. *Math. Ann.*, 289(2):315–334, 1991.

[20] Nimish A. Shah. Limit distributions of polynomial trajectories on homogeneous spaces. *Duke Math. J.*, 75(3):711–732, 1994.

[21] Nimish A. Shah. Limit distributions of expanding translates of certain orbits on homogeneous spaces. *Proc. Indian Acad. Sci. Math. Sci.*, 106(2):105–125, 1996.

[22] Nimish A. Shah. Counting integral matrices with a given characteristic polynomial. *Sankhyā Ser. A*, 62(3):386–412, 2000. Ergodic theory and harmonic analysis (Mumbai, 1999).

[23] Nimish A. Shah. Limiting distributions of curves under geodesic flow on hyperbolic manifolds. *Duke Math. J.*, 148(2):251-279, 2009.

[24] Nimish A. Shah. Asymptotic evolution of smooth curves under geodesic flow on hyperbolic manifolds. *Duke Math. J.*, 148(2):281-304, 2009.

[25] Nimish A. Shah. Equidistribution of expanding translates of curves and Dirichlets theorem on Diophantine approximation. *Invent. Math.*, 177(3):509-532, 2009.

[26] Nimish A. Shah. Expanding translates of curves and Dirichlet-Minkowski theorem on linear forms. *J. Amer. Math. Soc.*, 23:563–589, 2010.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Schur-Weyl Dualities and Link Homologies

Catharina Stroppel*

Abstract

In this note we describe a representation theoretic approach to functorial functor valued knot invariants with the focus on (categorified) Schur-Weyl dualities. Applications include categorified Reshetikhin-Turaev invariants, an extension of Khovanov homology and a diagrammatical description of the category of finite dimensional $GL(m|n)$-modules.

Mathematics Subject Classification (2010). Primary 17B10, 17B37, 57M27, 32S55

Keywords. Reshetikhin-Turaev invariants, knots, TQFT, general Lie supergroup, diagram algebras, Koszul algebras, 3j-symbols, Hecke algebra.

Introduction

The category $\mathcal{C}$ of finite dimensional modules over a complex semi-semisimple Lie algebra is a well-known semi-simple tensor category. A ground breaking generalization of this category appeared through the introduction of Quantum groups by Drinfeld and Jimbo ([Dr1], [Ji]), originally in the context of Yang-Baxter equations. In contrast to the category $\mathcal{C}$, the tensor category of finite dimensional modules over the corresponding quantum group comes along with a very interesting non-trivial braiding. Although at least generically still semi-simple, and therefore quite easy to handle, this braiding provides an important additional structure which was extensively used to construct knot invariants. The most basic one arising from the smallest quantum group $U_q(\mathfrak{sl}_2)$ is the Jones polynomial, or more general, the Reshetikhin-Turaev invariant [Jo1], [RT]. The first one was introduced by Jones in the 1980's in the context of von Neumann algebras (see [Jo1] for a wonderful overview), and is a (Laurent)-polynomial

*Mathematik Zentrum, Endenicher Allee 60, 53115 Bonn, Germany.
E-mail: stroppel@math.uni-bonn.de.

invariant of knots and links. The second one uses the representation theory of the quantum group in a much more subtle way and generalizes to invariants of tangles and 3-manifolds.

One crucial observation is that these structures have an integral version. In the 1990's Crane and Frenkel started to propagate the idea of categorifying integral structures. In this way one should be able to lift the above mentioned invariants to functorial valued invariants which should be finer and carry even more structure than the original ones. Crane and Frenkel presented an astonishing proposal for a possible invariant of 4-manifolds obtained via a partition function on the triangulation of the 4-manifold and conjectured that such an invariant can be brought into existence via some 'Hopf-categorification', i.e. the promoting of a Hopf algebra to an appropriate category. Although such a Hopf categorification has not yet been established (nor worked out axiomatically) there were several fundamental steps done in the last few years. One of the first successful categorifications in this context was obtained by Khovanov [K1] who categorified the Jones polynomial via some combinatorially defined categories. This so-called Khovanov homology turned out to be very powerful. One of the main applications is Rasmussen's combinatorial proof [Ra1] of the Milnor conjecture determining the slice genus of a (p, q)-torus knot. It is also known that Khovanov homology detects the unknot.

The most fascinating feature of Khovanov homology seems to me to be its connections to many different fields; there is no way one could mention all the applications, connections and occurrences of this homology theory. One of the basic problems is the following

Problem 1. *1. Construct a functorial Reshetikhin-Turaev tangle invariant extending Khovanov's categorification*

2. Construct a categorification of Reshetikhin-Turaev's 3*-manifold invariant*

A categorification of functorial Reshetikhin-Turaev tangle invariant was obtained by the author in 2003 ([St3]) which later was shown to agree with Khovanov's categorification of links and even tangles after restriction to a certain subcategory ([St4], [BS3]). The second part of the problem is much harder and so far open. The first step here should be a categorification of the tensor category $\mathrm{Rep}(U_q(\mathfrak{sl})_2)$ of finite dimensional representations of quantum $\mathfrak{sl}_2$ by defining a braided monoidal functor Ψ which assigns to an object $V_{d_1} \otimes V_{d_2} \otimes \cdots \otimes V_{d_r}$ in $\mathrm{Rep}(U_q(\mathfrak{sl})_2)$ a graded category, to a morphism between two objects an exact functor between the corresponding categories, and also lift the monoidal structure and the braiding. Then Clebsch-Gordon coefficients should have an interpretation in terms of dimensions of vector spaces, Jones-Wenzl projectors should become quotient functors etc.

In this article we want to indicate a representation theoretic approach to this task, where the associated categories are certain highest weight categories of modules for various $\mathfrak{gl}(n, \mathbb{C})$'s. To avoid too many technicalities, we do not want to present the whole construction of the functor Ψ here, but only

consider tensor products of the natural 2-dimensional representation V, where we have a very nice extra structure given by the so-called *Schur-Weyl duality*. We indicate how Khovanov homology arises from a categorification of Schur-Weyl duality, providing a natural explanation why this homology theory categorifies the Jones polynomial, and connecting it with highest weight Lie theory and categories of perverse sheaves. The main goal of this paper is to illustrate which important role is played by Schur-Weyl dualities in recent categorifications, constructions of link homologies and higher representation theory.

What are direct applications to representation theory? We obtain a combinatorial, elementary description of blocks of the parabolic category $\mathcal{O}$ for maximal parabolics in type A, as well as for blocks of finite dimensional representation for the Lie supergroup $\mathrm{GL}(m|n)$.

Outline of the paper: We start by recalling briefly the very classical setup of Schur-Weyl duality going back to the early 20th century, to work of Frobenius, Schur and Weyl. It connects the representation theory of the general linear group with that of the symmetric group. Based on this we explain the basic construction of the Reshetikhin-Turaev-Jones invariants for tangles using the quantum group $U_q(\mathfrak{sl}_2)$. Then the first theorem provides a categorification of the Schur-Weyl duality with resulting functor valued functorial knot invariants (which finally provides the above mentioned braided monoidal 2-functor Φ from the category of tangles to a certain category with objects certain derived categories). Theorem 2.5 provides the link between our categorification of the invariants using highest weight categories of representations of the general linear Lie algebra, Braden's description of the category of perverse sheaves on Grassmannians and the combinatorially defined Khovanov homology. The latter appears then naturally as a categorification of the space of $U_q(\mathfrak{sl}_2)$-invariant vectors inside our categorification.

We believe that our extra structure and information will be a key tool in the construction of 3-manifold invariants or for connecting the symplectic Khovanov homology with the original Khovanov theory. Some ideas are outlined below.

The second part of the paper relies on a *Schur-Weyl duality for higher levels*. This duality connects modules for $\mathfrak{gl}_n$ with modules over a cyclotomic version of Drinfeld's degenerate affine Hecke algebra (pioneered by Arakawa-Suzuki [AS] and Brundan-Kleshchev [BK1]). There are two main results here: the first one is a new proof of Theorem 2.5 (bypassing geometry completely) and constructing an interesting 2-Kac-Moody representation in the sense of [Ro] in complete detail, the second one is an interesting grading and a new presentation of the above mentioned cyclotomic quotient of Drinfeld's algebra in case of level 2 arising naturally from the Koszul grading on the category $\mathcal{O}$ (which should also be true for general level). These algebras turned up recently as a special case in independent work of Khovanov-Lauda, Rouquier and Vasserot-Varagnolo who

constructed algebraically (resp. geometrically) categorifications of the negative part of quantum groups. They have the potential to categorify Schur-Weyl dualities in general, and then provide a categorification of knot invariants from quantum groups of arbitrary semi-simple complex Lie algebras (see Section 2.4). These algebras also give a new insight into the representation theory of Hecke algebras.

The last part of the article is built on a super version of the higher Schur-Weyl duality. There we consider the category $\mathcal{F}(m|n)$ of finite dimensional modules over the Lie supergroup $GL(m|n)$. By results of Serganova [Ser] and Brundan [B1], the character formulas for simple modules are known with a given -more or less satisfactory- algorithm. It was observed a long time ago that these character formulas can be presented using Kazhdan-Lusztig polynomials for maximal parabolic in type A (i.e. the Grassmannian case). We make this precise and relate $\mathcal{F}(m|n)$ with modules over a generalized Khovanov algebra. In this way we obtain as a byproduct very easy formulas for the characters in terms of diagrams. It turns out that the category $\mathcal{F}(m|n)$ is actually equivalent to a certain limit version of the category of perverse sheaves on Grassmannians. This result might in fact replace the missing geometry (in form of a localization theorem) in this context.

Acknowledgement. I would like to thank H. Andersen, J. Brundan, S. Cautis, C. Haug, V. Mazorchuk, J. Sussan, I. Smith, and P. Teichner for various helpful discussions and comments on a previous version of this paper. I am particularly grateful to my math teacher Capo for his constant support.

1. Classical and Quantum Schur-Weyl Duality

For a fixed natural number k let $V = \mathbb{C}^k$ be the natural vector representation of $G = \mathrm{GL}(k, \mathbb{C})$. The symmetric group S_n acts on the tensor product $V^{\otimes n}$ by permuting the factors, obviously commuting with the G-action. The *Schur-Weyl duality* states that the subalgebras of $\mathrm{End}_{\mathbb{C}}(V^{\otimes n})$ generated by the image of the two actions are precisely each others' commutants, in particular all G-endomorphisms can be expressed in terms of the symmetric group. The image of the G-action is the *Schur algebra* $S(k, n)$. Then, if $k \geq n$, tensoring with the above (G, S_n)-bimodule $V^{\otimes n}$ defines an equivalence from the category of finite dimensional $S(k, n)$-modules (that means polynomial representations of G, homogeneous of degree n) to the category of finite dimensional S_n-modules. Instead of G one might prefer to work with the semi-simple Lie algebra $\mathfrak{sl}(k, \mathbb{C})$ or, equivalently, its universal enveloping algebra $\mathcal{U}(k) = \mathcal{U}(\mathfrak{sl}(k, \mathbb{C}))$. In the quantum Schur-Weyl duality ([Ji]) this picture gets then deformed: $\mathcal{U}(k)$ is replaced by the quantum group $\mathcal{U}_q(k)$ a certain Hopf algebra deformation of $\mathcal{U}(k)$ which in an appropriate way specializes to $\mathcal{U}(k)$. It acts now on $V = \mathbb{C}(q)^k$, and the group algebra of S_n gets replaced by a

q-deformation, the (generic) Hecke algebra $\mathcal{H}(S_n)$ over $\mathbb{C}(q)$. A simple transposition s_i does not act by an involution anymore. The action arises from an interesting braiding on the category $\mathrm{Rep}(U_q(k))$ of finite dimensional $\mathcal{U}_q(k)$-modules (the universal R-matrix). Again, the images of the two actions are each other commutants.

All the statements so far have an integral version ([Do], [Lu]). In the following we will tacitly use the $\mathbb{Z}[q,q^{-1}]$-form of $U_q(k)$ and $\mathcal{H}(S_n)$, but stick to the old notation.

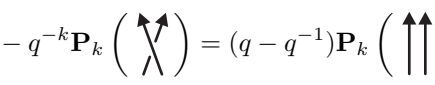
1.1. Invariants of tangles. In the basic case $k=2$, the image TL_n of the Hecke algebra action is called *Temperley-Lieb algebra* and easily explained in terms of the tensor structure of $\mathrm{Rep}(U_q(k))$. Fixing an isomorphism $V \cong V^*$ there are the (co)evaluation morphisms $\cup : \mathbb{C}(q) \to V \otimes V$ and $\cap : V \otimes V \to \mathbb{C}(q)$, and the image of the $\mathcal{H}(S_n)$-action is generated by the $\theta_i := 1^{\otimes(i-1)} \otimes \cup \circ \cap \otimes 1^{\otimes(n-i-1)} = \cap_i \circ \cup_i$, for $1 \le i \le n-1$ (deforming the elements $1+s_i \in \mathbb{C}[S_n]$). One can identify a $\mathbb{Z}[q,q^{-1}]$-basis of TL_n with isotopy classes of (n,n)-tangle diagrams with no crossings and no internal circles, such that the multiplication is given by concatenation of diagrams and replacing each circle by a scalar $q+q^{-1}$. For instance $\theta_i^2 = (q+q^{-1})\theta_i$ (deforming the equality $(1+s_i)^2 = 2(1+s_i)$ in $\mathbb{C}[S_n]$). We might reformulate the duality as

$$S_q(2,d) \curvearrowright V^{\otimes d} \curvearrowleft \mathrm{TL}_n \tag{1}$$

where $S_q(2,d)$ denotes the quantized Schur algebra. The faithful TL_n action identifies (n,n)-tangles without crossings and internal circles with basis vectors in the space of intertwiners of $V^{\otimes n}$. More generally, each (n,n')-tangle diagram t without crossings defines a $U_q(2)$-module homomorphism $P_2(t)$ from $V^{\otimes d}$ to $V^{\otimes d'}$ providing a bijection between isotopy classes of tangles with no crossings and internal circles and a $\mathbb{Z}[q,q^{-1}]$-basis of intertwiners. In other words we get a fully faithful functor P_2 from the Temperley-Lieb category with objects natural numbers and morphisms isotopy classes of tangle diagrams without crossing and no internal circles, to the subcategory $\mathcal{C}$ of $\mathrm{Rep}(U_q(2))$ with objects the various $V^{\otimes d}$ and morphisms generated by the $\cup_i$'s, and $\cap_i$'s.

1.2. Skein relations and crossings. The braid group action mentioned above associates to the crossings (displayed in (2)) of the i^{th} and $(i+1)^{th}$ strand the $\mathbb{Z}[q,q^{-1}]$-linear maps $q\theta_i - \mathrm{id}$ and $q^{-1}\theta_i - \mathrm{id}$ respectively. In this way, P_2 extends to the *Reshetikhin-Turaev invariant* of tangles [RT], with the skein relation

$$q^k \mathbf{P}_k\left(\;\right) - q^{-k}\mathbf{P}_k\left(\;\right) = (q-q^{-1})\mathbf{P}_k\left(\;\right) \tag{2}$$



where $k=2$. Any $(0,0)$-tangle is hereby mapped to an endomorphism of $\mathbb{Z}[q,q^{-1}]$, which is the multiplication by the *Jones* (Laurent)-*polynomial* ([Jo1]).

2. Categorification and Functorial Knot Invariants

A categorification of the classical Schur-Weyl-duality for $k = 2$ was suggested by Bernstein, Frenkel and Khovanov and completed in the quantum case in [St3], [FKS], see [MS2] for the general case. Involved here are certain (depending on k) categories of $\mathfrak{g} = \mathfrak{gl}(n, \mathbb{C})$-modules introduced in [BGG] and generalized in [R-C]. For $k = 2$, let $\mathcal{C}(n) = \bigoplus_{i=0}^{n} \mathcal{O}^{i,n-i}$ be the direct sum of $(i, n-i)$-parabolic subcategories of the principal block of the highest weight category $\mathcal{O}$ for $\mathfrak{g}$ equipped with the Koszul grading from [BGS]. Let $\mathcal{C}(n)^! = \bigoplus_{i=0}^{n} \mathcal{O}_{i,n-i}$ be the Koszul or quadratic dual category of $\mathcal{C}(n)$ given by certain singular blocks of $\mathcal{O}$. The Koszul grading turns the Grothendieck groups $\mathbb{K}_0(\mathcal{C}(n))$ and $\mathbb{K}_0(\mathcal{C}(n)^!)$ into $\mathbb{Z}[q, q^{-1}]$-modules, isomorphic to $V^{\otimes n}$. Important here is that each $\mathcal{O}^{i,n-i}$ is equivalent to the category of finite dimensional modules over some complex finite dimensional algebra $A^{i,n-1}$ which can naturally be equipped with a $\mathbb{Z}$-grading. It has $\binom{n}{i}$ isomorphism classes of simple modules, hence is suitable for categorifying a $\binom{n}{i}$-dimensional weight space of $V^{\otimes n}$.

Example 2.1. We have $A^{0,2} = A^{2,0} = \mathbb{C}$, whereas $A^{1,1}$ is isomorphic to the path algebra A of $\bullet \overset{1}{\underset{2}{\leftrightarrows}} \bullet$ with the relation $1 \to 2 \to 1$ being zero, and the grading given by path length. The intertwiner θ_i can be lifted to the functor $Ae_2A\otimes_{A}-$, where e_2 is the second primitive idempotent. Lifting the quantum group action involves passage to the derived category. On the other hand viewing Ae_2 as an $(A, \mathbb{C})$-bimodule defines (via tensoring) a functor which together with its adjoint can be used to lift the quantum group action. However to construct a commuting lift of θ_i one has again to pass to the derived category. The involved derived functors can be defined by saying that the two constructions are connected by Koszul duality (note that A is isomorphic to its quadratic dual $A^!$ in this special example).

The above example generalizes Lie theoretically to two categorifications of the quantum Schur-Weyl duality (1) (linked via Koszul duality):

$$\begin{array}{ccccc} \text{graded versions of certain} & \curvearrowright & D^b(\mathcal{C}(n)) & \curvearrowleft & \text{graded versions of certain} \\ \text{derived Zuckerman functors} & & & & \text{exact projective functors} \\ \text{graded versions of certain} & \curvearrowright & D^b(\mathcal{C}(n)^!) & \curvearrowleft & \text{graded versions of certain} \\ \text{exact projective functors} & & & & \text{derived Zuckerman functors} \end{array}$$

Note that in the first example the Temperley-Lieb algebra action is categorified via exact functors, the quantum group action however only exists when passing to the derived category. It is vice versa in the Koszul dual situation. To explain this in more detail note that the tensor category of finite dimensional $\mathfrak{g}$-modules acts via exact endofunctors $E \mapsto \_ \otimes E$ on $\mathcal{O}$. By the famous classification theorem of [BG], the endofunctors of the principal block of $\mathcal{O}$ obtained in this way form an additive tensor category with indecomposable

objects indexed by elements of the symmetric group, categorifying the action of the regular representation of the symmetric group. By definition, these functors restrict to endofunctors on each parabolic $\mathcal{O}^{i,n-i}$. It is still a mystery how these so-called *projective functors* decompose and behave under restriction to general parabolic $\mathcal{O}$'s. A crucial result of [St3] proves that they behave well under restrictions to $\mathcal{C}(n)$ via $F \mapsto \oplus F_{|\mathcal{O}^{i,d-i}}$. Graded versions (as defined in [St1]) of these projective functors restrict to an additive category with split Grothendieck ring isomorphic to TL_n acting on $\mathbb{K}_0(D^b(\mathcal{C}(n)))$ as desired. Under Koszul duality the functors become derived graded Zuckerman functors ([R-H], [MOS]). The quantum group action is given by a family of graded Zuckerman respectively projective functors naturally commuting with the TL_n-action. Koszul duality interchanges in some sense the two sides of the Schur-Weyl duality.

The main result of [St3] with [St4] is then the following:

Theorem 2.2. *1. The categorification of* TL_n *via graded projective functors extends to a functorial tangle and knot invariant which assigns to each* (n, n')*-tangle diagram a functor from* $D^b(\mathcal{C}(n))$ *to* $D^b(\mathcal{C}(n'))$ *inducing the above Reshetikhin-Turaev-Jones invariant* P_2 *on* $\mathbb{K}_0$*.*

2. This extends further to an invariant of cobordisms, well-defined up to scalars; each cobordism is sent to a natural transformation homogeneous of degree equal to the negative of the Euler characteristic of the cobordism.

By introducing certain markings ("disorientation lines") on cobordisms, (see [CMW]), the above construction finally defines a 2-functor from the 2-category with objects the natural numbers, morphisms tangle diagrams, and 2-morphisms cobordisms with disorientation lines into a category where objects are the categories $D^b(\mathcal{C}(n))$, morphisms are (certain) triangulated functors and 2-morphisms are (certain) natural transformations.

The $U_q(2)$-weight space decomposition corresponds to a decomposition into indecomposable abelian categories; the isotypic component decomposition only corresponds to a filtration of the categories (in the singular case by the Gelfand-Kirillov dimension, in the parabolic case by the annihilator, see [MS2]). The unique irreducible $(n+1)$-dimensional $U_q(2)$-summand V_n corresponds to a category $\mathcal{C}_n$ equivalent to

$$\bigoplus_{i=0}^{d} H^*(\mathrm{Gr}(i,d))\text{−gmod} \tag{3}$$

where $\mathrm{Gr}(i,d)$ denotes the Grassmannian of i-planes in $\mathbb{C}^d$. The $U_q(2)$-action is given by correspondences, see [FKS], passing between the direct summands.

An alternative categorification of $V^{\otimes n}$ was constructed in [CK1] using derived categories of (equivariant) coherent sheaves on a compactification of a resolution of the Slodowy slice to an adjoint orbit. The action of the category

of tangles is provided by certain explicit Fourier-Mukai transforms. There, the weight spaces do not correspond to direct summands. Conjecturally, the abelian categories arising from Lie theory are equivalent to subcategories of certain exotic t-structures (in the sense of [Be]), see [SW] for a more precise conjecture.

The above construction generalizes to arbitrary k. To get a functorial invariant satisfying (2) ones needs apart from Schur-Weyl duality a categorification of the tensor products of fundamental representations $\wedge^k V$ of $\mathfrak{gl}_n$. Intertwiners correspond to colored trivalent graphs satisfying the Murakami-Ohtsuki-Yamada relations. Such a categorification was established using $\mathcal{O}$ in [Su] and [MS3], and using coherent sheaves in [CK2]. These invariants suffer from the problem that, as given, they are quite hard to compute, but on the positive side provide natural situations for interesting braid group actions and carry many aspects of the integral representation theory of the original quantum group.

A different approach to functor valued knot invariants using (homotopy) categories of matrix factorizations was already developed in [KR1], [KR2]. We refer to [MS3] for an indication of a possible connection to the theories above.

2.1. Arbitrary tensor products. Built on (3), a powerful axiomatic theory of abelian categorifications of irreducible $\mathfrak{sl}_2$-modules was invented by Chuang and Rouquier [CR] and substantially further developed in [Ro] in form of 2-Kac-Moody representations. The higher structure rigidifies enough to obtain a unique categorification for each irreducible module. As advocated in [CF] such a 2-representation theory should provide a machinery that produces new categories out of some given categories, in particular interesting tensor categories. One of the challenging problems here is the following:

Problem 2. *Develop a 2-representation theory for tensor categories arising from quantum groups or more general Kac-Moody algebras.*

The answer should in particular include the existing abelian categorifications of arbitrary tensor products in $\mathrm{Rep}(\mathcal{U}_q(n))$ from [FKS], where $V_{d_1} \otimes \cdots \otimes V_{d_r}$ is categorified using Harish-Chandra bimodules with central character corresponding to $(d_1, d_2, \ldots, d_r)$, a quotient category of the above categorification of $V^{\otimes d}$, $d = \sum_{i=1}^{r} d_i$. This construction allows categorifications of the Jones-Wenzl projectors, the *colored* Reshetikhin-Turaev tangle invariant and $3j$-symbols, extending the original work of Khovanov [K2] in a new direction.

Conjecture 2.3. *There are renormalized $6j$-symbols which can be categorified using Harish-Chandra bimodules.*

In this way we hope to provide a first step in direction categorifying 3-manifold invariants. Details of this current work will appear in [FSS].

2.2. Braid group action and Serre functor. The Hecke algebra action on $\mathbb{K}_0(C(n))$ arises from a braid group action on $D^b(C(n))$ which is known to be faithful on each summand ([KS]). These braid group actions have

a long history in the representation theory of complex semisimple Lie algebras, known as *Enright-Joseph's completion*, *Irving's shuffling*, *Arkhipov's twisting* functors etc. and were originally introduced to study the so-called Kazhdan-Lusztig conjecture [KL], now a theorem, describing multiplicity formulas of simple composition factors of Verma modules. It is a well-supported principle that for any suitable braid group action on a category, the Serre functor will be given by the functor $C_{w_0}^2$ corresponding to the full (positive) twist w_0^2 in the center of the braid group, indeed ([MS1]):

Theorem 2.4. *Up to a shift (depending on i) in the derived category, the functor $C_{w_0}^2$ is the Serre functor of $D^b(\mathcal{O}^{i,n-i})$. Its square root C_{w_0} is (up to a shift) the Ringel duality functor.*

The quasi-hereditaryness of $\mathcal{O}^{i,n-i}$ implies that the categories are quite far away from being Calabi-Yau categories. However let $P^{i,n-i}$ be a minimal projective generator of $\mathcal{O}^{i,n-i}$ with endomorphism ring $A^{i,n-i}$ and consider the unique direct summand fixed under the Serre functor. Its endomorphism ring $E^{i,n-i}$ is then a symmetric algebra, see [MS1] for a more general statement. It is this algebra $E^{i,n-i}$, naturally arising from the Serre functor, which gives a precise connection to Khovanov homology as we explain now.

2.3. Khovanov homology. A combinatorially defined functorial knot invariant categorifying the Jones polynomial was constructed by Khovanov in [K1] with an extension to even tangles in [K3]. The resulting doubly graded homological invariant is called *Khovanov homology*. The following result describes the connection between our tangle invariant and Khovanov's:

Theorem 2.5. *Khovanov's arc algebra H_n is isomorphic as a graded algebra to the algebra $E^{n,n}$. Under this isomorphism, the combinatorially defined functor valued invariants are then obtained from the representation theoretically defined functor valued invariants by restriction.*

The isomorphism was first proved in [St4] using the fact that $\mathcal{O}^{i,n-i}$ is, via localization theorem and Riemann-Hilbert correspondence, equivalent to the category of perverse sheaves, constructible with respect to the Schubert stratification, on $\mathrm{Gr}(i,n)$, identifying Braden's explicit description of $A^{i,n-i}$ ([Bra]) with a combinatorially defined generalized Khovanov algebra. A second quite recent proof (bypassing geometry completely) with the explicit identification of the functor valued invariants was obtained in [BS3] and will be explained in more detail below. As predicted (see [K4]), the two tangle invariants have the same information, since the category $\mathcal{O}^{n,n}$ can be reconstructed from H_n via some double centralizer property, meaning that there is a functor from $A^{i,n-i}-\mathrm{mod}$ to $E^{i,n-i}-\mathrm{mod}$ fully faithful on projectives (see [St2] for a general statement). This property generalizes Soergel's structure theorem [So1] describing singular blocks of $\mathcal{O}$ by modules over the cohomology ring $H^*(G/P)$

of the associated partial flag variety, with the commutative Frobenius algebra $H^*(G/P)$ replaced by a (non-commutative) symmetric algebra.

Remark 2.6. Theorem 2.5 gives a tool to prove a refinement of Theorem 2.2: The Temperley-Lieb category comes equipped with a natural tensor structure given by composing horizontally: the tensor product on objects is just the sum, on morphisms it is given by putting the tangle diagrams next to each other. Then Theorem 2.2 extends to a (weak) tensor functor. Using also braid diagrams and their categorifications, it extends to a functor of braided tensor categories.

A construction of a singly graded knot homology theory in terms of Lagrangian intersection Floer homology of certain Stein varieties (more precisely the generic fibre of the adjoint quotient map for $\mathfrak{sl}(2n,\mathbb{C})$ restricted to a transversal slice of the nilpotent orbit of a nilpotent matrix of Jordan type (n,n)) was worked out by Seidel and Smith in [SS], see [Ma] for a conjectural realization in terms of Hilbert schemes. A categorified Schur-Weyl duality in this context is not yet available, and the precise relationship to Khovanov homology is still unclear. Conjecturally, enlarging the Lagrangian Floer homology by adding additional non-compact Lagrangians should provide the ring $A^{n,n}$, in analogy to Remark 3.1. Then to set up a fully faithful functor connecting the two theories and proving formality, passing from Khovanov's algebra to the Koszul algebra $A^{n,n}$ might be helpful, since the Hochschild cohomology is finite dimensional, and it might be possible to control higher A_∞-structures. In the case of the simplest algebra $A^{1,n-1}$ in our family of algebras, this is worked out explicitly in [Sei] (the case of the Milnor fibres of simple singularities of type $\mathbf{A_n}$). Let $A = A^{i,n-i}$ be one of our Koszul algebras from above. Let $K^\bullet$ be the Koszul resolution of the semi-simple degree zero part A_0 with the grading shifted such that the differentials are homogeneous of degree 1. Then the Hochschild cohomology $\mathbb{H}^*(A) = \bigoplus_{s,t} \mathbb{H}^s(A)_t$ is naturally bigraded such that $\mathbb{H}^s(A)_t$ is a subquotient of the space of degree t homogeneous maps inside $\mathrm{Hom}_{A-A}(K^s \otimes_{A^0} A, A)$. The space $\mathbb{H}^2(A) = \bigoplus_s \mathbb{H}^s(A)_{2-s}$ controls A_∞-deformations. Based on explicit calculations we strongly believe the following

Conjecture 2.7. $\mathbb{H}^s(A)_{2-s} = 0$ *if* $s \neq 0$, *in particular* $\mathbb{H}^2(A) = Z(A)_2$.

Here, $Z(A) = Z(A^{i,n-i})$ denotes the center of $A^{i,n-i}$. This is known to be canonically isomorphic to the cohomology ring of the corresponding $(i,n-i)$-Springer fibre (see the special case [St4, Theorem 4.5.2] of the general theorem from [B2], [St4]). Hence $Z(A)_2$ is $(n-1)$-dimensional.

2.4. Knot invariants for other types.

Recently, B. Webster [W] announced an amazing generalization of the above categorifications for arbitrary finite dimensional complex semi-simple Lie algebras based on Khovanov-Lauda's graphical calculus from [KLa].

3. Higher Schur-Weyl Duality and 2-representations

Higher Schur-Weyl duality relates the category $\mathcal{O}(\mathfrak{g})$ for $\mathfrak{g} = \mathfrak{gl}_n$ to cyclotomic quotients of the degenerate affine Hecke algebra H_d introduced by Drinfeld [Dr2]. H_d is the associative algebra which equals as a vector space $\mathbb{C}[x_1, \dots, x_d] \otimes \mathbb{C}S_d$. Multiplication is defined so that under the obvious inclusions $\mathbb{C}[x_1, \dots, x_d]$ and $\mathbb{C}S_d$ become subalgebras of H_d, together with the relations

$$s_i x_j = x_j s_i \text{ if } i \neq j, j+1, \qquad s_i x_{i+1} = x_i s_i + 1.$$

Let M be an arbitrary $\mathfrak{gl}_n$-module, then by [AS] the S_d-action on $V^{\otimes d}$ can be extended to an H_d-action on $M \otimes V^{\otimes d}$ such that x_1 acts by multiplication with the Casimir element on the first two factors $M \otimes V$. This defines commuting actions

$$\mathfrak{gl}_n \curvearrowright M \otimes V^{\otimes d} \curvearrowleft \mathrm{H}_d \tag{4}$$

hence a functor from the category $\mathcal{O}$ for $\mathfrak{gl}_n$ to modules over H_d. Of course, the image of either of the two actions depends on the choice of M. With an appropriate choice this defines a *higher Schur-Weyl duality*, see [BK1]. We now want to indicate two applications, first the proof of Theorem 2.5 and secondly a description of the category of finite dimensional $GL(i|j)$-modules for arbitrary i, j. The way how Schur-Weyl duality enters here is different from the way it entered Theorem 2.2: here the Hecke algebra action will arise as 2-morphisms in a categorification, whereas there it was given by functors. We present the main idea of the proof (following [BS3]) here, since we believe that this approach provides a quite general machinery to prove equivalences of categories without having a candidate of a functor available. To set up a connection between at the first sight totally unrelated categories, we first explain how they categorify certain $U_q(\mathfrak{gl}_\infty)$-modules.

3.1. Step 1: categorifications of certain $\mathfrak{gl}_\infty$-modules.

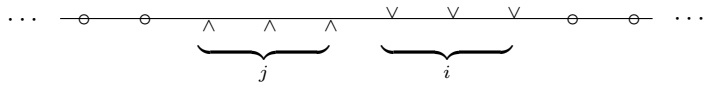
3.1.1. The Lie theory side. Fix non-negative integers i, $j = n - i$ and consider $\mathcal{O}(i, j)$, the sum of *all* integral blocks of the (i, j)-parabolic category $\mathcal{O}$ for $\mathfrak{gl}_n = \mathfrak{gl}_{i+j}$, equipped with the Koszul grading. Under the usual identification of integral weights of $\mathfrak{gl}_{i+j}$ with $\mathbb{Z}^{i+j}$, the simple objects in $\mathcal{O}(i, j)$ are (up to grading shifts) precisely the irreducible modules $L(\lambda)$ of highest weight $\lambda \in \Lambda(i, j)$, where $\Lambda(i, j) \subset \mathbb{Z}^{i+j}$ (after the usual shift with $\rho = (0, -1, -2, \dots, -(n-1))$) consists of tuples which are strictly decreasing in the first i entries as well as in the last j entries. Let $\overline{W} = \oplus_{s \in \mathbb{Z}} \mathbb{C} v_s$ denote the natural $U(\mathfrak{gl}_\infty)$-module (of infinite column vectors) and let W be its (integral) quantum version, then we have an isomorphism of $\mathbb{Z}[q, q^{-1}]$-modules

$$\Phi: \quad \mathbb{K}_0(\mathcal{O}(i, j)) \cong \bigwedge^i W \otimes \bigwedge^j W. \tag{5}$$

- If we choose Φ to send isomorphism classes of standard graded lifts of parabolic Verma modules to the standard basis, then simple modules are mapped to Lusztig's dual canonical basis, whereas the canonical basis corresponds to tilting modules (i.e. indecomposable projective modules twisted by the square root of the Serre functor). There are explicit formulas for the transformation matrices, based on [LS], [FK], easily expressible in terms of diagrams which motivated our construction of generalized Khovanov algebras (see below).

- Graded versions of projective functors categorify the $U_q(\mathfrak{gl}_\infty)$-action: there are functors $E_s, F_s : \mathcal{O}(i,j) \to \mathcal{O}(i,j)$, $s \in \mathbb{Z}$ lifting the action of the Chevalley generators e_s, f_s. The functor $F = \oplus_{s\in\mathbb{Z}} F_s$ is a suitably chosen graded version of tensoring with the natural $\mathfrak{gl}_n$-module V. (One might ask what categorifying means in this context. For our purposes it is enough to require that the linear maps on $\mathbb{K}_0$ induced by the functors E_s, F_s satisfy the Chevalley relations, in reality however we construct much more, namely a 2-Kac Moody representation in the sense of [Ro], see [BS3, Remark 5.7]).

3.1.2. The diagrammatical side. To each block Γ of $\mathcal{O}(i,j)$, we associate now a finite dimensional graded algebra K_Γ defined diagrammatically (generalizing Khovanov's arc algebra from [K1]). Each basis vector in $\bigwedge^i W \otimes \bigwedge^j W$ gets identified with a combinatorial weight in the sense of [BS3], i.e. with the diagram consisting of a number line whose vertices are indexed by $\mathbb{Z}$ and where the sth vertex is labeled $\vee$, $\wedge$, $\times$, $\circ$ depending on whether v_s occurs in the first, second, both or no tensor factor. Under this identification the isomorphism class of the parabolic Verma module of highest weight 0 for instance corresponds to

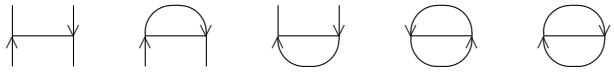


where the $\wedge$'s and $\vee$'s are on the vertices indexed $1-n, \ldots, -1, 0$. Two basis vectors correspond to the same block if they only differ by a permutation of $\wedge$' and $\vee$'s not touching the other labels. (In the above example there are $\binom{n}{i}$ basis vectors corresponding to the block Γ.) The algebra K_Γ has a vector space basis

$$\{(a\lambda b) \mid \text{for all oriented circle diagrams } a\lambda b \text{ with } L(\lambda) \in \Gamma\}.$$

given by triples (a, λ, b) of a cup diagram (involving cups and vertical rays), a combinatorial weight λ, and a cap diagram (involving caps and vertical rays) with some compatibility conditions. Its multiplication is defined by an explicit combinatorial procedure in terms of such diagrams (see [St4] for an alternative construction using a generalized 2-dimensional TQFT). For instance the

principal block of $\mathcal{O}(1,1)$ would correspond to an algebra with basis



This basis is homogeneous with grading given just by the number of clockwise cups and clockwise caps, in this case 0, 1, 1, 0, 2. The algebra structure is built such that it becomes isomorphic to the algebra $A^{1,1}$ from Example 2.1.

Remark 3.1. The diagrams for $\mathcal{O}^{i,n-i}$ from Section 2 have a natural interpretation in the theory of Springer fibres associated with 2-block nilpotent matrices of Jordan type $(i, n-i)$, indicating a direct connection to [CK2], [SS]. Weights naturally correspond to fixed points under a $\mathbb{C}^*$-action, the occurring cup diagrams correspond to the closures of fixed point attracting sets, the arcs indicate the type of flags they contain. Then our basis should be seen as labeling precisely triples $\{(x, L_1, L_2) \mid x \in L_1 \cap L_2\}$ of fixed points in the closure of pairwise intersections of two attracting sets. The graded vector space underlying our algebra is isomorphic to

$$\bigoplus_{(L_1,L_2)} H^*(L_1 \cap L_2)\langle \dim L_1 - \dim(L_1 \cap L_2)\rangle$$

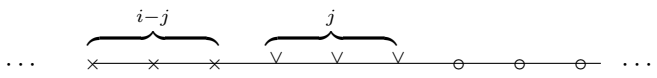
with the algebra structure given by a certain convolution product, see [SW]. We also want to mention that putting 1 $\wedge$ and $n-1$ $\vee$'s on arbitrary $n+1$ fixed vertices produces an algebra studied by Khovanov and Seidel [KS], [Sei].

Let K_Γ−gmod be the category of finite dimensional graded K_Γ-modules. Taking their direct sum over all blocks Γ of $\mathcal{O}(i,j)$ defines a category with the same properties as in Section 3.1.1. The action of the Chevalley generators is given by explicitly (graphically) defined bimodules. We also want to stress that the transformation matrix between the canonical and dual canonical basis already determines the dimension of the algebra. The construction in terms of triples (a, λ, b) should indicate the BGG-reciprocity formula passing between three bases of $\mathbb{K}_0$. One can show, [BS2], (purely combinatorially)

Theorem 3.2. *The algebra K_Λ is a graded Koszul quasi-hereditary algebra.*

3.2. Step 2: Higher structure: cyclotomic Hecke algebras.

3.2.1. Semisimple categories. For simplicity assume $i \geq n - i = j$. Each simple module $L(\lambda)$ in $\mathcal{O}$ with highest weight of the form



is a unique simple in its block, the same holds for its counter-part $L(\lambda)^{\mathrm{diag}}$ on the diagrammatical side. (Note that the only cup/cap diagram which could be put underneath or above to be oriented is the one containing rays only.) The corresponding blocks Λ are semi-simple, in particular equivalent. Under the isomorphism Φ these blocks correspond to highest weight vectors of $\bigwedge^i W \otimes \bigwedge^j W$.

3.2.2. Creating interesting categories from semisimple ones. The principal idea is now to construct two 2-categories, from $\mathcal{O}(i,j)$ and from the diagram side: objects are projective objects in the original category, morphisms are compositions of the functors categorifying the $U_q(\mathfrak{gl}_\infty)$ generators, their finite direct sums and finite summands, and 2-morphisms are natural transformations. Applying 1-morphisms to the semi-simple categories from the last section, one can create enough self-dual projective objects in either of the two categories. Finally one shows that higher Schur-Weyl duality provides enough natural transformations to control the endomorphism ring of a self-dual projective generator on either side and invokes a double centralizer property (see Section 2) to deduce an equivalence.

Let $L(\lambda)$ be as in the preceding section. Applying the higher Schur-Weyl duality [BK1] to $T := L(\lambda) \otimes V^{\otimes d} = F^d(L(\lambda))$ gives the first part of the following theorem, the others are more involved (see [BS3] for details)

Theorem 3.3. *1. The H_d-action on T factors through the cyclotomic quotient $H_d(i,j) := H_d/((x_1 - i)(x_1 - j))$ of level 2, inducing a surjective morphism of algebras $H_d(i,j) \to \mathrm{End}_{\mathcal{O}}(T)$.*

2. Let $T^{\mathrm{diag}} = F^d(L(\lambda)^{\mathrm{diag}})$. Then there is a homomorphism $H_d(i,j) \to \mathrm{End}(F^d)$ of algebras which induces under evaluation a natural surjection onto the endomorphism algebra of T^{diag}.

3. Via the above morphisms, the endomorphism rings of the projections of T resp. T^{diag} to a fixed block Γ, are both isomorphic to $e_\alpha H_d(i,j) e_\alpha$ for some appropriately chosen idempotent $e_\alpha \in H_d(i,j)$ depending on Γ. The composition of isomorphisms identifies the grading induced by the Koszul grading on $\mathcal{O}(i,j)$ with the diagrammatical grading.

Using the combinatorics from step 1 and a double centralizer construction one can then deduce Theorem 2.5. An interesting direct consequence is the following

Corollary 3.4. *The algebra $R^\Lambda_\alpha := e_\alpha H_d(i,j) e_\alpha$ inherits a $\mathbb{Z}$-grading from the grading on $\mathcal{O}(i,j)$ respectively the naive grading on the diagram algebras.*

4. A Graded Presentation of Cyclotomic Blocks

Corollary 3.4 predicts a quite unusual presentation of the level 2-quotients of Drinfeld's degenerate affine Hecke algebra compatible with the grading. Using the diagrammatically defined algebra this can be made completely explicit as follows: Under (5), a block is contained in a single weight space whose weight differs from the weight Λ obtained from λ by subtraction of a positive root $\alpha = \alpha_{i_1} + \alpha_{i_2} + \cdots + \alpha_{i_d}$ of height d. Let $\mathbf{I}^\alpha$ denote the S_d orbit of $(i_1, i_2, \ldots, i_d)$. Then there is a presentation of R^Λ_α where generators are

$$\{e(\boldsymbol{i}) \mid \boldsymbol{i} \in I^\alpha\} \cup \{y_1, \ldots, y_d\} \cup \{\psi_1, \ldots, \psi_{d-1}\},$$

with relations

$$y_1^{(\alpha_{i_1}, \Lambda)} e(\boldsymbol{i}) = 0; \quad e(\boldsymbol{i})e(\boldsymbol{j}) = \delta_{\boldsymbol{i},\boldsymbol{j}} e(\boldsymbol{i}); \quad \textstyle\sum_{\boldsymbol{i} \in I^\alpha} e(\boldsymbol{i}) = 1;$$
$$y_r e(\boldsymbol{i}) = e(\boldsymbol{i}) y_r; \quad \psi_r e(\boldsymbol{i}) = e(s_r \cdot \boldsymbol{i})\psi_r; \quad y_r y_s = y_s y_r;$$
$$\psi_r y_s = y_s \psi_r \quad \text{if } s \neq r, r+1; \quad \psi_r \psi_s = \psi_s \psi_r \quad \text{if } |r-s| > 1;$$

$$\psi_r y_{r+1} e(\boldsymbol{i}) = \begin{cases} (y_r \psi_r + 1) e(\boldsymbol{i}) & \text{if } i_r = i_{r+1}, \\ y_r \psi_r e(\boldsymbol{i}) & \text{if } i_r \neq i_{r+1}; \end{cases}$$

$$y_{r+1} \psi_r e(\boldsymbol{i}) = \begin{cases} (\psi_r y_r + 1) e(\boldsymbol{i}) & \text{if } i_r = i_{r+1}, \\ \psi_r y_r e(\boldsymbol{i}) & \text{if } i_r \neq i_{r+1}; \end{cases}$$

$$\psi_r^2 e(\boldsymbol{i}) = \begin{cases} 0 & \text{if } i_r = i_{r+1}, \\ (i_{r+1} - i_r)(y_{r+1} - y_r) e(\boldsymbol{i}) & \text{if } i_r = i_{r+1} \pm 1, \\ e(\boldsymbol{i}) & \text{otherwise;} \end{cases}$$

$$\psi_r \psi_{r+1} \psi_r e(\boldsymbol{i}) = \begin{cases} (\psi_{r+1}\psi_r\psi_{r+1} + (i_{r+1} - i_r)) e(\boldsymbol{i}) & \text{if } i_{r+2} = i_r = i_{r+1} \pm 1, \\ \psi_{r+1}\psi_r\psi_{r+1} e(\boldsymbol{i}) & \text{otherwise.} \end{cases}$$

By inspecting the relations it follows that there is a $\mathbb{Z}$-grading on R^Λ_α defined by declaring the e's to be of degree 0, the y's is of degree 2, and $\psi_r e(\boldsymbol{i})$ of degree $-2, 1$ or 0 according to whether $i_r = i_{r+1}$, $|i_r - i_{r+1}| = 1$ or $|i_r - i_{r+1}| > 1$. This is precisely the grading inherited from $\mathcal{O}$. This statement should be true in general, not only for maximal parabolic blocks of category $\mathcal{O}$.

4.1. Khovanov-Lauda-Rouquier-Varagnolo-Vasserot algebras. The above algebra turns out to be isomorphic to a level two cyclotomic quotient of an algebra associated with the Dynkin quiver of type A^∞, denoted $R(\alpha; \Lambda)$ in [KLa], and arising in a family of algebras constructed (independently) algebraically by Khovanov-Lauda and Rouquier, and geometrically by Vasserot-Varagnolo ([KLa], [Ro], [VV]). These algebras were introduced to categorify the negative part of quantum groups. Our approach gives a conceptual interpretation of the somehow (at least in the algebraic definition) artificial

looking grading on R^Λ_α. Although the algebras R^Λ_α are not quasi-hereditary, they have the nice structure of a graded cellular algebra in the sense of [GL]. Our methods yield a special *graded cellular basis* for R^Λ_α parameterized by some diagrams which are in bijection with certain Young tableaux, see [BKW] where the existence of such bases is predicted. In particular we deduce from this a graded dimension formula for the irreducible R^Λ_α-modules (in level two for finite type A). The construction of a graded cellular basis was generalized to higher levels in [HM].

5. $GL(m|n)$-modules Via Super Higher Schur-Weyl

The principal idea of the proof of Theorem 2.5 indicated above can also be applied (in a super version) to the category of finite dimensional integrable modules for the Lie superalgebra $\mathfrak{gl}(m|n)$, i.e. blocks for the Lie supergroup $GL(m|n)$, see [BS4] for details. Here m, n are arbitrary positive integers (playing **not** the same role as in the previous sections). The main result here is

Theorem 5.1. *Let $\mathbb{C}$ be a fixed algebraically closed field of characteristic 0.*

1. *Any block of $GL(m|n)$ of atypicality r is Morita equivalent to H^∞_r, a certain algebra (usually infinite dimensional) arising as a limit of generalized Khovanov algebras (built from combinatorial weights with r $\vee$'s and infinitely many $\wedge$'s).*

2. *These algebras are symmetric, quasi-hereditary and Koszul.*

Note that Koszulity is proved by completely elementary means using the diagram algebras (see Theorem 3.2). In the case of the super group we cannot invoke geometry, since so far no satisfying localization theorem is available. The theorem suggest that the missing geometry might not necessarily be found in the world of super flag varieties, but rather as a limit version of the categories of perverse sheaves on ordinary Grassmannians. (The diagrammatic approach gives as a byproduct a complete elementary proof of the Koszulity for $\mathcal{O}(i, j)$ without passing to perverse sheaves).

5.1. The category of finite dimensional $GL(m|n)$-modules.

To explain more details fix again $m, n \geq 0$ and let G denote the algebraic supergroup $GL(m|n)$ over $\mathbb{C}$, that is the functor from the category of commutative superalgebras over $\mathbb{C}$ to the category of groups, mapping a commutative superalgebra $A = A^{\bar{0}} \oplus A^{\bar{1}}$ to the group $G(A)$ of all invertible $(m + n) \times (m + n)$ matrices of the form

$$g = \left(\begin{array}{c|c} a & b \\ \hline c & d \end{array}\right) \tag{6}$$

where a (resp. d) is an $m \times m$ (resp. $n \times n$) matrix with entries in $A^{\bar{0}}$, and b (resp. c) is an $m \times n$ (resp. $n \times m$) matrix with entries in $A^{\bar{1}}$.

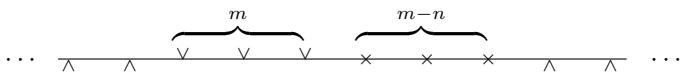
We are interested here in finite dimensional representations of G equivalently in integrable supermodules over its Lie superalgebra $\mathfrak{gl}(m|n, \mathbb{C})$. Allowing only even G-morphisms between G-modules turns it into an abelian category which decomposes into blocks. We pick one from each equivalence class under parity change and denote the resulting category $\mathcal{F}(m|n)$. The simple objects are then in bijection with dominant weights $X^+(T)$ for the standard torus T and Borel B.

By [B1], the category $\mathcal{F}(m|n)$ is a highest weight category. In analogy to (3.1.1) we obtain an isomorphism (only of $\mathbb{Z}$-modules, since there is no grading available)

$$\mathbb{K}_0(\mathcal{F}(m|n)) \cong \bigwedge^m \overline{W} \otimes \bigwedge^n \overline{W}^\star. \tag{7}$$

As in case of category $\mathcal{O}$ we have simple, indecomposable tilting, and standard or Verma modules (usually called *Kac modules* after [Ka]) giving rise to three distinguished bases.

5.2. The diagrammatics. Now we turn again our attention to the diagram algebra side and identify $X^+(T)$ with the set $\Lambda^{\text{super}} = \Lambda(m|n)$ of all diagrammatical weights with a total of m vertices labeled $\times$ or $\vee$, a total of n vertices labeled $\circ$ or $\vee$, and all of the (infinitely many) remaining vertices are labeled $\wedge$. The ("super version" of the) identification rule is now different from before: the i-th vertex is labelled $\times$, $\circ$, $\vee$, $\wedge$ depending on whether v_i occurs in the first tensor factor, in the second, in both, or in none. For example, assuming $m \geq n$, the zero weight parameterizing the trivial G-module is now identified with the diagram



where the leftmost $\vee$ is on vertex $(1 - m)$. Blocks usually have now infinitely many simple objects. The usual notion of *atypicality* in the representation theory of $GL(m|n)$ as in e.g. [Ser] is here just the number of $\vee$'s. Atypicality zero means the category is semi-simple. In terms of the corresponding diagrammatical algebra it is half the top degree. The construction of the diagram algebras works fine in this more general context, but produces infinite dimensional non-unital algebras. Theorem 3.2 is still valid. (For a general treatment of infinite dimensional Koszul algebras see [MOS]).

5.3. The equivalence. Let $\mathcal{K}(m|n)$ denote the direct sum of the module categories for the diagram algebras K_Γ, $\Gamma \subset \Lambda^{\text{super}}$.

Theorem 5.2. *There is an equivalence of highest weight categories*

$$\mathbb{E}: \quad \mathcal{F}(m|n) \to \mathcal{K}(m|n).$$

Consequences.

- In the diagrammatic setting the following non-trivial result of Serganova from [Ser] becomes obvious: the blocks of $GL(m|n)$ for all m, n depend up to equivalence only on the degree of atypicality of the block (not on m, n).
- Blocks of $GL(m|n)$ are Koszul, in particular can be equipped with a grading.
- When combined with the results from [BS3], our results can be used to prove the *Super Duality Conjecture* as formulated in [CWZ]. A direct algebraic proof of this conjecture, and its substantial generalization from [CW], has recently been found by Cheng and Lam [CL].

All of these results suggest some more direct geometric connection between the representation theory of $GL(m|n)$ and the category of perverse sheaves on Grassmannians may exist. The above result gives a very concrete and explicit description of the category $\mathcal{F}(m|n)$, but unfortunately not well adapted to the tensor product structure on this category. It is a challenge to find a categorification of the Schur-Weyl duality for tensor products of the natural representation for $U_q(\mathfrak{gl}(1|1))$ with a result similar to Theorem 2.5. In this way one should be able to solve the following

Problem 3. *Find an algebro-representation theoretic categorification of the Alexander polynomial P_0 from a categorification of the representation theory of $\mathfrak{gl}(1|1)$.*

Note that the Alexander polynomial P_0 is the Euler characteristic of a bigraded knot homology theory, discovered by Ozsvath-Szabo [OS] and Rasmussen [Ra1]. A categorification of a (super or not super) higher Schur-Weyl duality analogous to Theorem 2.5 is (so far) not available.

References

[AS] T. Arakawa and T. Suzuki, *Duality between $\mathfrak{sl}_n(\mathbb{C})$ and the degenerate affine Hecke algebra*, J. Algebra **209** (1998), 288–304.

[BG] J. Bernstein, S. Gelfand, *Tensor products of finite- and infinite-dimensional representations of semisimple Lie algebras*, Compositio Math. **41** (1980), no. 2, 245–285.

[BGG] J. Bernstein, I. Gelfand, S. Gelfand, *A certain category of $\mathfrak{g}$-modules*, Funkcional. Anal. i Prilozen. **10** (1976), no. 2, 1–8.

[BFK] J. Bernstein, I. Frenkel, M. Khovanov, *A categorification of the Temperley-Lieb algebra and Schur quotients of $U(\mathfrak{sl}_2)$ via projective and Zuckerman functors*, Selecta Math. (N.S.) **5** (1999), no. 2, 199–241.

[Be] R. Bezrukavnikov, *Quasi-exceptional sets and equivariant coherent sheaves on the nilpotent cone*, Represent. Theory **7** (2003), 1–18.

[Bra] T. Braden, *Perverse sheaves on Grassmannians*, Canad. J. Math. **54** (2002), no. 3, 493–532.

[B1] J. Brundan, *Kazhdan-Lusztig polynomials and character formulae for the Lie superalgebra* $\mathfrak{gl}(m|n)$, J. Amer. Math. Soc. **16** (2003), 185–231.

[B2] J. Brundan, *Symmetric functions, parabolic category* $\mathcal{O}$*, and the Springer fiber*, Duke Math. J. **143** (2008), no. 1, 41–79.

[BK1] J. Brundan and A. Kleshchev, *Schur-Weyl duality for higher levels*, Selecta Math. **14** (2008), 1–57.

[BK2] J. Brundan and A. Kleshchev, *Blocks of cyclotomic Hecke algebras and Khovanov-Lauda algebras*, to appear in Invent. Math.; `arXiv:0808.2032`.

[BK5] J. Brundan and A. Kleshchev, *Graded decomposition numbers for cyclotomic Hecke algebras*f, to appear in Adv. Math.; `arXiv:0901.4450`.

[BKW] J. Brundan, A. Kleshchev and W. Wang, *Graded Specht modules*. `arXiv:0901.0218`.

[BS2] J. Brundan, C. Stroppel: *Highest weight categories arising from Khovanov's diagram algebra II: Koszulity*, Transf. groups, **15** (2010), no.1, 1–45.

[BS3] J. Brundan, C. Stroppel: *Highest weight categories arising from Khovanov's diagram algebra III: Category* $\mathcal{O}$, `arXiv:0812.1090`.

[BS4] J. Brundan, C. Stroppel: *Highest weight categories arising from Khovanov's diagram algebra IV: the general linear supergroup*, `arXiv:0907.2543.`

[CL] S.-J. Cheng and N. Lam, *Irreducible characters of general linear superalgebra and super duality*, `arXiv:0905.0332`.

[CW] S.-J. Cheng and W. Wang, *Brundan-Kazhdan-Lusztig and super duality conjectures*, Publ. Res. Inst. Math. Sci. **44** (2008), 1219–1272.

[CWZ] S.-J. Cheng, W. Wang and R.B. Zhang, *Super duality and Kazhdan-Lusztig polynomials*, Trans. Amer. Math. Soc. **360** (2008), 5883–5924.

[CF] L. Crane, I. Frenkel: *Four dimensional topological quantum field theory, Hopf categories, and the canonical bases*, J. Math. Phys. **35** (1994), no. 10, 5136–5154.

[CMW] D. Clark, S. Morrison, K. Walker: *Fixing the functoriality of Khovanov homology*, Geom. Topol. **13** (2009), no. 3, 1499–1582

[CK1] S. Cautis, K. Kamnitzer, *Knot homology via derived categories of coherent sheaves. I. The* $\mathfrak{sl}(2)$*-case*, Duke Math. J. **142** (2008), no. 3, 511–588

[CK2] S. Cautis, K. Kamnitzer, *Knot homology via derived categories of coherent sheaves. II.* $\mathfrak{sl}_m$ *case*, Invent. Math. **174** (2008), no. 1, 165–232.

[CR] J. Chuang, R. Rouquier, *Derived equivalences for symmetric groups and* $\mathfrak{sl}_2$*-categorification*, Ann. of Math. (2) **167** (2008), no. 1, 245–298.

[BGS] A. Beilinson, V. Ginzburg, W. Soergel, *Koszul duality patterns in representation theory*, J. Amer. Math. Soc. **9** (1996), no. 2, 473–527.

[Do] S. Donkin, *On Schur algebras and related algebras I*, J. Algebra **104** (1986), 310-328.

[Dr1] V. Drinfeld, *Quantum groups*. Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986), 798–820.

[Dr2] V. Drinfeld, *Degenerate affine Hecke algebras and Yangians*, Func. Anal. Appl. **20** (1986), 56–58.

[FK] I. Frenkel and M. Khovanov, *Canonical bases in tensor products and graphical calculus for $U_q(\mathfrak{sl}_2)$*, Duke Math. J. **87** (1997), 409–480.

[FKS] I. B. Frenkel, M. Khovanov and C. Stroppel, *A categorification of finite-dimensional irreducible representations of quantum $sl(2)$ and their tensor products*, Selecta Math. (N.S.) **12** (2006), no. 3–4, 379–431.

[FSS] I. Frenkel, C. Stroppel and J. Sussan, *Tangle invariants, Clebsch-Gordon coefficients and 6j-symbols*, in preparation.

[GL] J. Graham and G. Lehrer, *Cellular algebras*, *Invent. Math.* **123** (1996), 1–34.

[HM] J. Hu, A. Mathas, *Graded cellular bases for the cyclotomic Khovanov-Lauda-Rouquier algebras of type A*, `arXiv:0907.2985`.

[Ji] M. Jimbo, *A q-analogue of $U(\mathfrak{gl}(N+1))$, Hecke algebra, and the Yang-Baxter equation*, Lett. Math. Phys., **10** (1985), 63–69.

[Jo1] V. Jones, *A polynomial invariant for knots via von Neumann algebras*, Bull. Amer. Math. Soc. (N.S.) Volume **12**, Number 1 (1985), 103–111.

[Jo2] V. Jones, V. Jones, *von Neumann algebras in mathematics and physics*. A plenary address presented at the International Congress of Mathematicians held in Kyoto, August 1990. ICM-90.

[Ka] V. Kac, *Characters of typical representations of classical Lie superalgebras*, *Comm. Algebra* **5** (1977), 889–897.

[KL] D. Kazhdan, G. Lusztig, *Representations of Coxeter groups and Hecke algebras*. Invent. Math. **53** (1979), no. 2, 165–184.

[K1] M. Khovanov, *A categorification of the Jones polynomial*, Duke Math. J. **101** (2000), no. 3, 359–426.

[K2] M. Khovanov, *Categorifications of the colored Jones polynomial*, J. Knot Theory Ramifications 14 (2005), no. 1, 111–130.

[K3] M. Khovanov, *A functor-valued invariant of tangles*, Algebr. Geom. Topol. 2 (2002), 665–741.

[K4] M. Khovanov, *Link homology and categorification*, International Congress of Mathematicians. Vol. II, 989–999, Eur. Math. Soc., Zürich, 2006.

[KLa] M. Khovanov and A. Lauda, *A diagrammatic approach to categorification of quantum groups I*, Represent. Theory 13 (2009), 309–347.

[KR1] M. Khovanov, L. Rozansky, *Matrix factorizations and link homology*, Fund. Math. **199** (2008), no. 1, 1–91.

[KR2] M. Khovanov, L. Rozansky, *Matrix factorizations and link homology*, II. Geom. Topol. **12** (2008), no. 3, 1387–1425.

[KS] M. Khovanov, P. Seidel, *Quivers, Floer cohomology, and braid group actions*, J. Amer. Math. Soc. **15** (2002), no. 1, 203–271.

[LS] A. Lascoux and M.-P. Schützenberger, *Polynômes de Kazhdan et Lusztig pour les Grassmanniennes*, Astérisque **87–88** (1981), 249–266.

[Le] E. Lee, *An endomorphism of the Khovanov invariant*, Adv. Math. **197** (2005), no. 2, 554–586.

[Lu] G. Lusztig, *Introduction to Quantum Groups*, Birkhäuser Boston, 1993.

[Ma] C. Manolescu, *Link homology theories from symplectic geometry*, Adv. Math. **211** (2007), no. 1, 363–416.

[MOS] V. Mazorchuk, S. Ovsienko, C. Stroppel, *Quadratic duals, Koszul dual functors, and applications*, Trans. Amer. Math. Soc. **361** (2009), 1129–1172.

[MS1] V. Mazorchuk, C. Stroppel, *Projective-injective modules, Serre functors and symmetric algebras*, J. Reine Angew. Math. **616** (2008), 131–165.

[MS2] V. Mazorchuk, C. Stroppel, *Categorification of (induced) cell modules and the rough structure of generalised Verma modules*, Adv. Math. **219** (2008), no. 4, 1363–1426.

[MS3] V. Mazorchuk, C. Stroppel, *A combinatorial approach to functorial quantum $\mathfrak{sl}_k$ knot invariants*, Amer. J. Math. **131** (2009), 16791713.

[OS] P. Ozsvath, Z. Szabo, *Holomorphic disks, link invariants and the multivariable Alexander polynomial*, Algebr. Geom. Top. **8** (2008) 615692.

[Ra1] J. Rasmussen, *Floer homology and knot complements*, PhD Thesis, Harvard University, 2003.

[Ra2] J. Rasmussen, *Khovanov homology and the slice genus*, `arXiv:0402131`.

[RT] N. Reshetikhin, V. Turaev, *Ribbon graphs and their invariants derived from quantum groups*, Comm. Math. Phys. **127** (1990), 126.

[R-C] A. Rocha-Caridi, *Splitting criteria for $\mathfrak{g}$-modules induced from a parabolic and the Bernstein-Gelfand-Gelfand resolution of a finite-dimensional, irreducible $\mathfrak{g}$-module*, Trans. Amer. Math. Soc. **262** (1980), no. 2, 335–366.

[Ro] R. Rouquier, *2-Kac-Moody algebras*, `arXiv:0812.5023`.

[R-H] S. Ryom-Hansen, *Koszul duality of translation- and Zuckerman functors*, J. Lie Theory **14** (2004), no. 1, 151–163

[Sc] I. Schur, *Über die rationalen Darstellungen der allgemeinen Gruppe*, Sitzungsberichte Akad. Berlin 1927, 5875 (1927).

[Sei] P. Seidel, *Fukaya Categories and Picard-Lefschetz Theory*, Zürich Lectures in Advanced Mathematics, 2008.

[SS] P. Seidel, I. Smith, *A link invariant from the symplectic geometry of nilpotent slices*, Duke Math. J. **134** (2006), no. 3, 453–514.

[Ser] V. Serganova, *Characters of irreducible representations of simple Lie superalgebras*, Documenta Math., ICM 1998 volume II, 583–596.

[So1] W. Soergel, *Kategorie $\mathcal{O}$, perverse Garben und Moduln über den Koinvarianten zur Weylgruppe*. J. Amer. Math. Soc. **3** (1990), no. 2, 421–445.

[St1] C. Stroppel, *Category $\mathcal{O}$: gradings and translation functors*, J. Algebra **268** (2003), no. 1, 301–326.

[St2] C. Stroppel, *Category $\mathcal{O}$: quivers and endomorphism rings of projectives*, Represent. Theory **7** (2003), 322–345.

[St3] C. Stroppel, *Categorification of the Temperley-Lieb category, tangles, and cobordisms via projective functors*, Duke Math. J. **126** (2005), no. 3, 547–596.

[St4] C. Stroppel, *Parabolic category $\mathcal{O}$, perverse sheaves on Grassmannians, Springer fibres and Khovanov homology*, Compos. Math. **145** (2009), no. 4, 954–992.

[St4] C. Stroppel, *TQFT with corners and tilting functors in the KacMoody case*, `arXiv:0605103`.

[SW] C.Stroppel, B. Webster, *2-block Springer fibers: convolution algebras and coherent sheaves*, `arXiv:0802.1943`.

[Su] J. Sussan, *Category $\mathcal{O}$ and $\mathfrak{sl}(k)$ link invariants*, `arXiv:0701045`.

[VV] M. Varagnolo and E. Vasserot, *Canonical bases and Khovanov-Lauda algebras*; `arXiv:0901.3992`.

[W] B. Webster, *Knot invariants and higher representation theory*, to appear.

[W] H. Weyl, *The Classical Groups. Their Invariants and Representations*, Princeton University Press, Princeton, N.J., 1939.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Cohomology of Arithmetic Groups and Representations

T. N. Venkataramana*

Abstract

We give a survey of results on restriction of cohomology classes on locally symmetric spaces to smaller locally symmetric spaces; these results are closely connected with cohomological representations of semi-simple Lie groups associated with the locally symmetric spaces and we describe the connection.

Mathematics Subject Classification (2010). Primary 11F75; Secondary 22E40, 22E41

1. Introduction

If $S(\Gamma) = \Gamma\backslash X$ is an arithmetic quotient of a Hermitian symmetric domain X (a connected component of a " Shimura Variety") then a natural class of subvarieties that one can costruct explicitly are quotients of Hermitian subdomains by smaller arithmetic subgroups ("Shimura Subvarieties"). It is easy to see from the "homotopy version" of the Lefschetz hyperplane section theorem that these subvarieties are not intersections of hyperplane sections.

However, one may consider all the translates of these Shimura Subvarieties under Hecke operators and ask for (a cohomological version of) a weaker Lefschetz property for the collection of these Hecke translates.

In [Oda], it is shown that Hecke translates of the Jacobian of a fixed Shimura curve span the Albanese of a quotient of the unit ball in $\mathbb{C}^n$ by an arithmetic group of the group $SU(n,1)$ of automorphisms of the unit ball in $\mathbb{C}^n$. This proves a version of the Lefschetz Theorem on the injection of the cohomology to Shimura curves.

There are a number of criteria developed in recent years to determine if Hecke translates a given cohomology class on a Shimura Variety, restricts non-trivially to a given Shimura subvariety. We give a survey of these results. These results are formulated in terms of the "representation type ("$A_{\mathfrak{q}}$") to

*School of Mathematics, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay - 400 005, INDIA. E-mail: venky@math.tifr.res.in.

which the cohomology class belongs. The criteria can be extended even to non-hermitian cases, and are expressed in terms of the compact dual of the symmetric space under consideration.

2. Notation and Statements

Fix two semi-simple algebraic groups H and G defined over $\mathbb{Q}$ and a morphism $j : H \to G$ of algebraic groups defined over $\mathbb{Q}$, with finite kernel. Fix a maximal compact subgroup K_H of $H(\mathbb{R})$ and extend $j(K_H)$ to a maximal compact subgroup K_∞ of $G(\mathbb{R})$. We have then an embedding $j : X_H \to X_G$ of the symetric spaces $X_H = H(\mathbb{R})/K_H$ and $X_G = G(\mathbb{R})/K_\infty$.

If $\Gamma \subset G(\mathbb{Q})$ is a torsion-free congruence arithmetic group, then the quotient $S(\Gamma) = \Gamma\backslash X_G$ is a manifold covered by X_G. Denote by $\mathbb{A}_f$ the ring of finite adeles over $\mathbb{Q}$ and by $G(\mathbb{A}_f)$ the group of $\mathbb{A}_f$ rational points. The group $G(\mathbb{R})$ acts on X_G and $G(\mathbb{A}_f)$ acts on the left on $G(\mathbb{A}_f)$; hence $G(\mathbb{Q}) \subset G(\mathbb{R}) \times G(\mathbb{A}_f)$ acts diagonally on $X_G \times G(\mathbb{A}_f)$. Also, $G(\mathbb{A}_f)$ acts (by right multiplication on the second factor) on $X \times G(\mathbb{A}_f)$. Hence $G(\mathbb{A}_f)$ acts on the quotient $S_G = G(\mathbb{Q})\backslash X_G \times G(\mathbb{A}_f)$. Moreover, S_G is the inverse limit $S_G(K) = S_G/K$ where $K \subset G(\mathbb{A}_f)$ is a compact open subgroup. The space $S_G(K)$ is a finite union of locally symmetric manifolds $S(\Gamma)$ for a finite set of Γ.

Denote by $H^*(S_G)$ the cohomology of S_G with complex coefficients. Then (by [Rohlfs]), the cohomology ring $H^*(S_G)$ is the direct limit over $K \subset G(\mathbb{A}_f)$ of the cohomology groups $H^*(S_G(K), \mathbb{C})$ on which $G(\mathbb{A}_f)$ acts via its right action on S_G. If $g \in G(\mathbb{A}_f)$ and $\omega \in H^*(S_G)$, then we denote by $g^*(\omega)$ the action of g onω.

We have similarly the space $S_H = H(\mathbb{Q})\backslash X_H \times H(\mathbb{A}_f)$ and a map $j : S_H \to S_G$.

We can now define the "Oda restriction map" (see [Oda])

$$Res : H^*(S_G) \to \prod_{g \in G(\mathbb{A}_f)} H^*(S_H),$$

defined by $Res(\omega) = (j^* g^*(\omega))_{g \in G(\mathbb{A}_f)}$.

In this survey we are concerned with describing the kernel of Res in terms of representation theory.

If G is anisotropic over $\mathbb{Q}$, then S_G is compact and by the Matsushima formula we have the decomposition

$$H^*(S_G) = \oplus m(\pi) H^*(\mathfrak{g}, K_\infty, \pi_\infty) \otimes \pi_f.$$

In this formula, $\pi = \pi_\infty \otimes \pi_f$ is a representation of the group $G(\mathbb{A}) = G(\mathbb{R}) \times G(\mathbb{A}_f)$ which occurs in $L^2(G(\mathbb{Q})\backslash G(\mathbb{A})$ and π_∞ is a cohomological representation, i.e. the relative Lie algebra cohomology space $H^*(\mathfrak{g}, K_\infty, \pi_\infty) \neq 0$, where $\mathfrak{g}$ is the complexification of the Lie algebra of $G(\mathbb{R})$, and $m(\pi)$ is the

multiplicity of the representation $\pi = \pi_\infty \otimes \pi_f$ of $G(\mathbb{R}) \times G(\mathbb{A}_f) = G(\mathbb{A})$ in $L^2(G\mathbb{Q})\backslash G(\mathbb{A})$.

The representations with cohomology, of $G(\mathbb{R})$ are classified (by the work of Parthasarathy, Kumaresan, Vogan and Zuckerman) in terms of the θ-stable parabolic subalgebras $\mathfrak{q}$ of the complex semi-simple Lie algebra $\mathfrak{g}$, with θ being the Cartan involution on $G(\mathbb{R})$ with respect to the maximal compact subgroup K_∞. If $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ is the associated Cartan decomposition, we have the θ-stable Levi decomposition $\mathfrak{q} = \mathfrak{l} \oplus \mathfrak{u}$ of the parabolic subalgebra $\mathfrak{q}$ and the decomposition $\mathfrak{u} = \mathfrak{u} \cap \mathfrak{k} \oplus \mathfrak{u} \cap \mathfrak{p}$. Put $R = dim(\mathfrak{u} \cap \mathfrak{p})$.

The Cartan decompositon $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ is a decomposition of K_∞ modules. The line $\wedge^R(\mathfrak{u} \cap \mathfrak{p})$ generates an irreducible representation $V(\mathfrak{q})$ of K_∞ in $\wedge^R\mathfrak{p}$.

The classification of unitary irreducible cohomological representations π_∞ of $(\mathfrak{g}, K_\infty)$ now says that to each θ-stable parabolic subalgebra $\mathfrak{q}$ as above, there exists a cohomological representation $A_\mathfrak{q}$ characterised by the property that the only irreducible K_∞ representation common to $\wedge^*\mathfrak{p}$ and $A_\mathfrak{q}$ is the representation $V(\mathfrak{q})$. Moreover, every cohomological representation π_∞ is an $A_\mathfrak{q}$.

If $\omega \in H^R(S_G)$, and under the Matsushima decomposition, ω lies in the component $H^R(\mathfrak{g}, K_\infty, \pi_\infty) \otimes \pi_f m(\pi)$, where $\pi_\infty = A_\mathfrak{q}$ and $R = dim(\mathfrak{u} \cap \mathfrak{p})$, we will then refer to ω as a **strongly primitive** class of type $A_\mathfrak{q}$.

Denote by $\widehat{X_G}$ and $\widehat{X}_H$ the compact dual symmetric spaces of X_G and X_H. The Matsushima component corresponding to the trivial representation of $G(\mathbb{A})$ is siomorphic to $H^*(\widehat{X}_G)$. The submanifold $\widehat{X}_H$ yields a cohomology class (its fundamental class) in $H^*(\widehat{X}_G) \subset H^*(S_G)$, denoted $[\widehat{X}_H]$.

The Levi subgroup $L(\mathbb{C}) \subset Q(\mathbb{C}) \subset G(\mathbb{C})$ is defined over $\mathbb{R}$ and is θ-stable. We have an associated map of compact symmetric spaces $\widehat{X}_L \subset \widehat{X}_G$, and the restriction map $\widehat{Res} : H^*(\widehat{X}_G) \to H^*(\widehat{X}_L)$. We have then the following criterion for the non-vanishing of the Oda-restriction purely in terms of the compact dual of X_G ([V1]):

Theorem 1. *If ω is a strongly primitive cohomology class of type $A_\mathfrak{q}$ in $H^R(S_G)$, and if $\widehat{Res}([\widehat{X}_H]) \neq 0$ in $H^*(\widehat{X}_L)$, then the Oda restriction of ω is non-zero.*

As a corollary, we get the following result ([V1]) (conjectured by M.Harris and J-S.Li ([H-L]), and proved by them in degrees $i \leq 2$).

Theorem 2. *If $G(\mathbb{R}) = SU(n,1)$ and $H(\mathbb{R}) = SU(m,1)$ up to compact factors and $j : H \to G$ induces the standard embedding of $SU(m,1)$ in $SU(n,1)$, then the restriction map*

$$Res : H^i(S_G) \to \prod_{g \in G(\mathbb{A}_f)} H^i(S_H)$$

is injective for $i \leq m$.

The criterion of Theorem 1 is especially useful in the case when both X_G and X_H are Hermitian symmetric domains and the embedding $j : X_H \to$

X_G is holomorphic. Then we have the K_∞-equivariant decomposition of the complexified tangent space $\mathfrak{p}$ into holomorphic and anti-holomorphic tangent spaces $\mathfrak{p} = \mathfrak{p}^+ \oplus \mathfrak{p}^-$. Similarly, for the subalgebra $\mathfrak{h}$ we have $\mathfrak{h} = \mathfrak{h} \cap \mathfrak{k} \oplus \mathfrak{h} \cap \mathfrak{p}$, and $\mathfrak{p}_H = \mathfrak{h} \cap \mathfrak{p}$ decomposes into a direct sum of $\mathfrak{p}_H^+$ and $\mathfrak{p}_H^-$.

The embedding j induces a map $\mathfrak{p}_H^+ \to \mathfrak{p}^+$.

Moreover, $\mathfrak{u} \cap \mathfrak{p} = \mathfrak{u} \cap \mathfrak{p}^+ \oplus \mathfrak{u} \cap \mathfrak{p}^-$. We have $\mathfrak{g} = \mathfrak{u} \oplus \mathfrak{l} \oplus \mathfrak{u}^-$ where $\mathfrak{u}^-$ is the nilradical opposite to $\mathfrak{u}$ and stable under $\mathfrak{l}$.

Write $R^\pm = dim \mathfrak{u} \cap \mathfrak{p}^\pm$. Then $R = R^+ + R^-$. Set $V^+(\mathfrak{q})$=span of K_∞ translates of the line $\wedge^{R^+}(\mathfrak{u} \cap \mathfrak{p}^+) \wedge \wedge^{R^-}(\mathfrak{u}^- \cap \mathfrak{p}^+)$ in the K_∞-representation $\wedge^R \mathfrak{p}^+$.

Denote by $E(G, H, R)$ the K_∞-span of the subspace $\wedge^R \mathfrak{p}_H^+$. When X_G is Hermitian, note that S_G is a projective limit of algebraic varieties.

We have then the necessary condition ([V1]) for the non-vanishing of *Res*:

Theorem 3. *If G is anisotropic over $\mathbb{Q}$ and is of Hermitian type, and if ω is a strongly primitive class on S_G of Hodge type (R^+, R^-) and of type $A_\mathfrak{q}$, then $Res(\omega) \neq 0$ provided $V^+(\mathfrak{q}) \cap E(G, H, R) \neq 0$.*

When the class ω is of Hodge type $(m, 0)$ (i.e. $R = R^+$), this criterion is necessary *and* sufficient ([Cl-V]):

Theorem 4. *If G is anisotropic over $\mathbb{Q}$ and is of Hermitian type, and if ω is a holomorphic form of degree R on S_G of typ $A_\mathfrak{q}$, then $Res(\omega) \neq 0$ if and only if $E(G, H, R) \supset V(\mathfrak{q})$.*

The case when the cohomology classes are not of holomorphic type is more involved and this is the result of Theorem 3; however, in this case, the criterion of Theorem 3 is only proved to be sufficient.

2.1. Applications to cup-products.

We take $G/\mathbb{Q}$ as before, of Hermitian type. Replace the pair (H, G) by the diagonal embedding $(G, G \times G)$. The restriction to the diagonal G of a tensor product class $\omega_1 \otimes \omega_2 \in H^* S_G) \otimes H^*(S_G) = H^*(S_{G \times G})$ is simply the cup product, and from Theorem 3 we get

Theorem 5. *If ω_1 and ω_2 are two strongly primitive classes on S_G of type $A_{\mathfrak{q}_1}$ and $A_{\mathfrak{q}-2}$, then for some $g \in G(\mathbb{A}_f)$ the cup product $g^*(\omega_1) \wedge \omega_2 \neq 0$ if $V^+(\mathfrak{q}_1) \wedge V^+(\mathfrak{q}_2) \neq 0 \subset \wedge^* \mathfrak{p}^+$.*

In case the classes are holomorphic, this is actually necessary and sufficient thanks to a result of Clozel ([Clo 2]) In Parthasarathy ([Par]) a sufficient condition for the vanishing of cup products is given.

As an application of Theorem 5, we have([V1]): if $G/\mathbb{Q}$ is such that $G(\mathbb{R}) = SU(n, 1)$ up to compact factors, then given $\omega_1 \in H^i(S_G)$ and $\omega_2 \in H^j(S_G)$ (not necessarily primitive), the cup product $g^*(\omega_1) \wedge \omega_2 \neq 0$ for some $g \in G(\mathbb{A}_f)$. Analogous results were proved earlier by Kudla ([Ku]).

2.2. Cycles on Shimura Varieties. The results (Theorem 3 and Theorem 1) may be used to prove some results on cycles on compact Shimura varieties ([V2] and [V3]).

Let $G/\mathbb{Q}$ be an anisotropic semi-simple group such that $G(\mathbb{R})$ is, up to compact factors, isomorphic to $SU(n,2)$. It can be shown that $H^4(S_G)$ is a direct sum of $H^{4,0} \oplus H^{0,4}$ and $H^{2,2}$ as $\mathbb{Q}$-Hodge structures. Moreoer, we may write $H^{2,2}(S_G) = H^4(\widehat{X}_G) \oplus W$ where W consists of non-$G(\mathbb{A}_f)$ invariant classes. Using the foregoing criteria, one may prove that W restricts injectively into $H^2(SU(2,1)) \otimes H^2(SU(2,1)$; one may even prove that W restricts into a product of Hodge classes: $W \subset H^{1,1}_{\mathbb{Q}}(SU(1,2)) \otimes H^{1,1}_{\mathbb{Q}}(SU(1,2))$. Using the Lefschetz (1,1) Theorem, we now get ([V2])

Corollary 1. *All the Hodge classes in $H^{2n-2,2n-2}(S_G)$ are generated by $G(\mathbb{A}_f)$-translates of fundamental classes of products of curves i.e. classes of the form $[C_1 \times C_2]$ where C_1 and C_2 are curves and $C_1 \times C_2$ embeds in S_G/K for some compact open subgroup $K \subset G(\mathbb{A}_f)$.*

The criteria of Theorem 3 and Theorem 1 can be applied to prove non-triviality of certain cycle classes as well as the occurrence of certain cohomological representations in the automorphic spectrum. The following can be shown.

Corollary 2. *If $X_H \subset X_G$ is an embedding of Hermitian domains, there exists a hlomorphic cohomology class on S_G which restricts non-trivially to S_H and if the centraliser of $H(\mathbb{R})$ in $G(\mathbb{R})$ is strictly larger than the centre of G, then the $G(\mathbb{A}_f)$- module generated by the cycle class $[S_H(\Gamma)]$ is infinite dimensional.*

In particular, the existence of holomorphic automorphic representations $A_{\mathfrak{q}}$, implies the automorphy of $A_{\mathfrak{q}}$ with $A_{\mathfrak{q}}$ of Hodge type (p,p).

Examples: (1) If $G = U(p,q)$ and $H = \prod_{1 \leq i \leq r} U(p_i, q_i)$ with $\sum p_i = p$ or $\sum q_i = q$.

(2) $G = Sp_g$ and $H = Sp_{g_1} \times \cdots \times Sp_{g_r}$, with $\sum g_i = g$.

In contrast, if these equalities are not satisfied (i.e. $\sum p_i < p$ and $\sum q_i < q$ and $\sum g_i < g$, then the cycle class $[S_H(\Gamma)]$ generates the trivial $G(\mathbb{A}_f)$-module.

These and similar computations raise the possibility that the following may have a positive answer.

Question 1. Given a simple Lie group G defined over $\mathbb{Q}$ and a semi-simple $\mathbb{Q}$-subgroup H such that the centralizer of H in G is non-compact, is it always the case that $[S_H(\Gamma]$ lies in $H^*(\widehat{X}_G)$ (i.e. generates the trivial $G(\mathbb{A}_f)$-module)?

2.3. Mumford-Tate Groups. The conjectures of Langlands on the zeta functions on Shimura Varieties (and their extension to the non-tempered case by Kottwitz and Arthur) predict that in low degrees, the Galois group (of the number field over which a Shimura variety is defined) acts by a "small" group. In particular, for very low degrees of cohomology, the Galois action is potentially abelian. This is equivalent to saying (modulo the Mumford-Tate

conjecture on the relation between the Galois group and the Mumford-Tate group) that the Mumford-Tate group of the $\mathbb{Q}$-Hodge structure associated to low degree cohomology is abelian. This implication can be proved for several arithmetic quotients of classical Hermitian domains. (see [Bla-Rog], [Mu-Ra2] for related results).

Theorem 6. *(1) If $G(\mathbb{R}) = Sp_g$ and $g \geq 2$ then the Mumford-Tate group of the $\mathbb{Q}$-Hodge structure of $H^g(S_G)$ is abelian.*

(2) If $G(\mathbb{R}) = SU(p,q)$ and $2 \leq p \leq q$, then the Mumford-Tate group of the $\mathbb{Q}$-Hodge structure associated to $H^p(S_G)$ is abelian.

Here is a sketch of the proof. We use the criteria of restriction to deduce that the cohomology restricts injectively to a product of Shimura Curves in S_G, and then use the fact that the Hodge types of cohomological representations in low degrees are highly restricted (Vogan-Zuckerman). Then the following Lemma completes the proof.

Lemma 7. *Suppose that W is an irreducible pure $\mathbb{Q}$-Hodge structure whose Hodge types are holomorphic or anti-holomorphic: $W \otimes \mathbb{C} = W^{m,0} \oplus W^{0,m}$, with $m \geq 0$. Suppose that W is contained in a tensor product of two irreducible $\mathbb{Q}$-Hodge structures A and B, such that $A \otimes \mathbb{C} = \oplus_{p,q \geq 0, p+q=a} A^{p,q}$ and $B \otimes \mathbb{C} = \oplus_{p,q \geq 0, p+q=b} B^{p,q}$. Then the Mumford-Tate groups of W, A and B are all Abelian.*

3. The Action of the Cohomology of the Compact Dual

Under the Matsushima decomposition

$$H^*(S_G) = \oplus m(\pi) H^*(\mathfrak{g}, K_\infty, \pi_\infty) \otimes \pi_f,$$

the part which corresponds to the trivial representation π is the cohomology of the compact dual $H^*(\widehat{X_G})$. Therefore, it acts on the cohomology of S_G) by cup product. If X is Hermitian symmetric, it is possible to split the Hodge structure $H^*(S_G)$ into smaller pieces according to this action.

Suppose $\pi_\infty = A_\mathfrak{q}$ and $\pi'_\infty = A_{\mathfrak{q}'}$ are two cohomological representations which have strongly primitive cohomology in degree i. Suppose that L and L' are respectively the Levi subgroups of the parabolic subgroups Q and Q' corresponding to the θ-stable parabolic sublagebras $\mathfrak{q}$ and $\mathfrak{q}'$. We consider the restriction maps $r_L : H^*(\widehat{X_G} \to H^*(\widehat{X_L}$ and $r_{L'} : H^*(\widehat{X_G} \to H^*(\widehat{X_{L'}}$. Denote by $Hod^i(A_\mathfrak{q})$ the smallest Q-Hodge structure whose complex points contain all the strongly primitive cohomology classes in degree i of type $A_\mathfrak{q}$. Define similarly, $Hod^i(A_{\mathfrak{q}'})$.

Theorem 8. *If the kernels of the maps r_L and $r_{L'}$ are distinct, then the $\mathbb{Q}$ Hodge Structures $Hod^i(A_\mathfrak{q})$ and $Hod^i(A_{\mathfrak{q}'})$ are disjoint (their intersection is the zero vector space).*

As an example, consider G such that $G(\mathbb{R}) = SU(2,2)$ up to compact factors. Then, in degree $i = 2$, there are three parabolic subalgebras $\mathfrak{q}$ whose $A_{\mathfrak{q}}$ have cohomology in degree 2. Two of them (say $\mathfrak{q}_1$ and $\mathfrak{q}_2$ are holomorphic (of Hodge type $(2,0)$) and the other (say, $\mathfrak{q}_3$) is of Hodge type $(1,1)$. It can be verified from the criterion of Theorem 8 that the associated $\mathbb{Q}$-Hodge structures are all disjoint. By the Lefschetz $(1,1)$-theorem, the Hodge structure associated to $\mathfrak{q}_3$ consists of algebraic cycles. This can be shown to yield the following.

Corollary 3. *If $G(\mathbb{R}) = U(2,2)$ up to compact factors, then all Tate classes in $H^2(S_G)$ are algebraic.*

Remark 1. The Tate conjecture for H^2 for most Shimura varieties is known, in all dimensions at least five (by unpublished work of Blasius and Rogawski (a much earlier work of Harder-Langlands ([Har-Lan]) , Murty-Ramakrishnan ([Mu-Ra]) and Klingenberg ([Kli]) treats the case of Hilbert modular surfaces). The above Corollary shows that for $U(2,2)$ also, the Tate Conjecture holds. The main open case is then that of compact quotients of the two fold product of the upper half plane by cocompact irreducible lattices in $SL_2(\mathbb{R}) \times SL_2(\mathbb{R})$.

4. Non-Hermitian Case

To tackle the general (non-hermitian) case, M.Harris and J-S.Li devised an alternative approach ([H-L]). This is in terms of the "automorphic dual" of G in the sense of Burger, Li and Sarnak ([Bu-Sa]). Recall that $G(\mathbb{R})$ is a real semi-simple Lie group and denote by $\widehat{G(\mathbb{R})}$ the space of equivalence classes of irreducible unitary representations of $G(\mathbb{R})$ under the Fell topology (of uniform convergence of matrix coefficients of representations on compact subsets of $G(\mathbb{R})$). Denote by $\widehat{G(\mathbb{R})}_{Aut}$ the closure of the union (over all congruence subgroups Γ) of the collections of irreducibles π which occur weakly in $L^2(\Gamma\backslash G(\mathbb{R}))$.

The following conjecture is due to many people ([H-L], [Ber] and [Ber-Cl]).

Conjecture 1. (Harris-Li, Bergeron and Clozel) If π is a cohomological representation, then π is not a limit of complementary series representations σ with $\sigma \in \widehat{G(\mathbb{R})}_{Aut}$.

In particular, if π is a non-tempered cohomological representation, then it is isolated in the automorphic dual od G.

A very special case of this is when π is the trivial representation of SL_2 and the conjecture is equivalent to saying that the non-zero eigenvalues of the Laplacian on quotients of the upper half plane by congruence subgroups of $SL_2(\mathbb{Z})$ are bounded away from zero (Selberg's "3/16" Theorem ([Sel]). For a general semi-simple group G defined over $\mathbb{Q}$, Clozel has proved that conjecture 1 is true for the trivial representation ([Clo]). A result of Vogan ([Vog]) says that for most groups, the cohomological representations $A_{\mathfrak{q}}$ are isolated even in the unitary dual (the only ones which are not isolated are those for which the Levi subgroup L over $\mathbb{R}$ is a product of copies of $SO(m,1)$ or $SU(m,1)$).

Harris and Li showed that under the assumption of Conjecture 1, the question of the non-vanishing of the restriction of cohomology may be reduced purely to a question of the discrete occurrence of a suitable cohomological representation of the smaller group $H(\mathbb{R})$ in a cohomological representation of the larger group $G(\mathbb{R})$. In the special case that $G(\mathbb{R}) = SU(n,1)$ (up to compact factors), they proved that Shimura subvarieties of the complex hyperbolic manifold S_G satisfy a Lefschetz property namely

$$Res : H^i(S_G) \to \prod_{g \in G(\mathbb{A}_f)} H^* S_H),$$

is injective provided $i \leq dim S_H$ (they even proved this unconditionally in the case that $i \leq 2$).

Clozel and Bergeron have an analogue for the *real* hyperbolic manifolds under the assumption of Conjecture 1, namely the following theorem.

Theorem 9. *(Clozel and Bergeron). Under the assumption of Conjecture 1, if $G(\mathbb{R}) = SO(n,1)$ and $H(\mathbb{R}) = SO(m,1)$ up to compact factors, then the restriction map*

$$Res : H^i(S_G) \to \prod_{g \in G(\mathbb{A}_f)} H^i(S_H),$$

is injective provided $i \leq [m/2]$ (where $[x]$ is the integral part of x).

Clozel and Bergeron have shown that Conjecture 1 follows from well known conjectures of Arthur on the possible non-tempered automorpihc representations which can arise. Becuse of recent progress on the Fundamental Lemma, and results of Arthur on consequences of the Fundamental Lemma, Conjecture 1 is close to being settled.

Theorem 9 has been proved unconditionally (*only for* $i = 1$) by [Ra-V], [Lub] and [V4]).

In [Sp-V], Conjecture 1 for $SO(n,1)$ is reduced to the case when the cohomological representation is tempered.

Acknowledgments

I extend to M.S. Raghunathan, R. Parthasarathy, L.Clozel and M.V. Nori my hearty thanks for many very helpful conversations over the years on the material related to this survey.

References

[Ber] Bergeron, N. Lefschetz Properties of Arithmetic Real and Complex Hyperbolic Manifolds, IMRN (2003), **no. 20**, 1089–1122.

[Ber-Cl] Bergeron, N. and Clozel, L. Spectre Automorpphe de varietés de hyperboliques et applications de topologiques, Asterisque no 303 (2003).

[Bla-Rog] Blasius, D. and Rogawski, J. Zeta functions of Shimura Varieties, Motives (Seattle WA (1991), Proc. Sympos. Pure Math., **55**, Part 2, Amer Math. Soc., Providence, RI, 1994.

[Bo-Wa] Borel, A. and Wallach, N., Continuous Cohomology, Discrete Subgroups and Representations of real reductive groups, Princeton University Press, Princeton, N.J. (1980).

[BLS] Burger, M., Li, J-S. and Sarnak, P. Ramanujan Duals and automorphic spectrum, Bull.Amer.Math.Soc. (N.S.)**26** (1992), 253–257.

[Bu-Sa] Burger, M. and Sarnak, P. Ramanujan Duals II. Invent.Math.**106** (1991) 1–11.

[Cl-V] Clozel, L. and Venkataramana, T.N. Restriction of holomorphic cohomology of a Shimura variety to a smaller Shimura variety, Duke Math Journal,**98**, (1998), **no. 1**, 51–106.

[Clo] Clozel, L. Démonstration de la conjecture τ. Invent. Math.**151** (2003), 297–328.

[Clo 2] Clozel, L. Produits dans cohomologie holomorphe de varietés de Shimura, J Reine Angew Math.**430**, (1992) 69–83.

[Clo 3] Clozel, L. Produits dans cohomologie holomorphe varietes de Shimura II J.Reine Angew Math. **444**, (1993), 1–15.

[Har-Lan] Harder, G. and Langlands, R.P. Algebraische Zyklen auf Hilbert-Blumenthal Flachen, Journal Fur die Reine und Angew. Math., **366** (1986), 53–120.

[H-L] Harris, M. and Li, Jian-Shu, A. Lefschetz property for subvarieties of Shimura Varieties, J. Algebraic Geom., **7** (1998), **no. 1**, 77–122.

[Kli] Klingenberg C., Die Tate Vermutung fur Hilbert-Blumenthal Flachen, Invent.Math., **89**, (1987), 291–317.

[Ku] Kudla, S. Algebraic Cycles on Shimura Varieties of Orthogonal Type, Duke Math. J., **86** (1997) **no.1**, 39–78.

[Lub] Lubotzky, A. Eigenvalues of the Laplacian, the First Betti Number and the Congruence Subgroup Problem, Ann. of Math. **2 144**, (1996), no. 2 441–452.

[Mu-Ra] Murty, K. and Ramakrishnan, D. Period Relations and the Tate Conjecture for Hilbert Modular Surfaces, Invent. Math. **89** (1987), no. 2 319–345.

[Mu-Ra2] Murty, K. and Ramakrishnan, D. The Albanese of Unitary Shimura Varieties, The Zeta Functions of Picard-Modular Surfaces, 445–464, Univ. Montreal, Montreal, QC, 1992.

[Oda] Oda, T. A Note on the Albanese Variety of an arithmetic quotient of the complex hyperball, J.Fac. Sci. Univ. Tokyo Sect.IA Math. **28** (1981), no. 3, 481–486.

[Par] Parthasarathy, R. Holomorphic Forms on $\Gamma\backslash G/K$ and Chern Classes, Topology **21** (1982), no. 2, 151–178.

[Ra-V] Raghunathan, M.S. and Venkataramana, T.N. First Betti number and the congruence subgroup problem. Linear Algebraic Groups and Their Representations, Los Angeles, CA, 1992), 95–107, Contemp. Math. **153**, Amer. Math.Soc., Provdence RI, 1993.

[Rohlfs] Rohlfs, J., Projective Limits of Locally Symetric Spaces and Cohomology, J. Reine Angew Math. **479** (1996), 149–182.

[Sel] Selberg, A. On the Estimation of Fourier Coefficients of modular forms, Proc. Symp Pure Math. Amer Math Soc. vol 8 (1965), 1–15.

[Sp-V] Speh, B. and Venkataramana, T.N. Discrete Components of some Complementary Series, (to appear) Forum Mathematicum, (2010).

[V1] Venkataramana, T.N. Cohomology of Compact Locally Symmetric Spaces, Compositio Math 2000.

[V2] Venkataramana, T.N. On Cycles on Compact Locally Symmetric Varieties, Monats Math. **135** (2002) **no. 3**, 221–244.

[V3] Venkataramana, T.N. Some Remarks on Cycles on Compact Shimura Varieties, J. Ramanujan Math. Soc., **16**, (2001), **no. 4** 309–322.

[V4] Venkataramana, T.N. Restriction Maps and the First Betti number, Algebraic Groups and Arithmetic, 91–97, Tata Institute of Fundamental Research, Mumbai 2004.

[V-Z] Vogan, D. and Zuckerman, G. Unitary Representations with nonzero cohomology, Compositio Math. **53**, (1984), **no.1**, 51–90.

[Vog] Vogan, D. Isolated Unitary Representations, Automorphic Forms and Applications, 379–398, IAS/Park City Math. Ser., **12**, AMS, Providence, RI, (2007).

Section 8

Analysis

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Differentiability of Lipschitz Functions, Structure of Null Sets, and Other Problems

Giovanni Alberti, Marianna Csörnyei*, and David Preiss

Abstract

The research presented here developed from rather mysterious observations, originally made by the authors independently and in different circumstances, that Lebesgue null sets may have uniquely defined tangent directions that are still seen even if the set is much enlarged (but still kept Lebesgue null). This phenomenon appeared, for example, in the rank-one property of derivatives of BV functions and, perhaps in its most striking form, in attempts to decide whether Rademacher's theorem on differentiability of Lipschitz functions may be strengthened or not.

We describe the non-differentiability sets of Lipschitz functions on $\mathbb{R}^n$ and use this description to explain the development of the ideas and various approaches to the definition of the tangent fields to null sets. We also indicate connections to other current results, including results related to the study of structure of sets of small measure, and present some of the main remaining open problems.

Mathematics Subject Classification (2010). Primary 26B05; Secondary 28A75.

Keywords. Lipschitz, derivative, tangent, width, unrectifiability

1. Differentiability of Lipschitz Functions

One of the important results of Lebesgue tells us that Lipschitz functions on the real line are differentiable almost everywhere. It is also well-known that the converse is true: for every Lebesgue null set E on the real line there is a real-valued Lipschitz function which is non-differentiable at any point of E. That is:

*Department of Mathematics, University College London, Gower Street, London, WC1E 6BT, United Kingdom. E-mail: mari@math.ucl.ac.uk.

Theorem 1.1. *For a given set $E \subset \mathbb{R}$ there is a Lipschitz function $f\colon \mathbb{R} \to \mathbb{R}$ which is not differentiable at any point $x \in E$ if and only if E is Lebesgue null.*

One of our aims is to generalise Theorem 1.1, and also its more precise variants that will be described in Theorem 1.13, to Lipschitz functions $f\colon \mathbb{R}^n \to \mathbb{R}^m$.

Since a Lipschitz function on $\mathbb{R}$ is differentiable almost everywhere, Fubini Theorem implies immediately that the *directional (or partial) derivative*

$$f'(x;u) := \lim_{t\to 0} \frac{f(x+tu)-f(x)}{t}$$

of a Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ exists for each direction u at a.e. x.

Although differentiability is not the same as the existence of sufficiently many partial derivatives, the set of points at which these two notions differ is relatively easy to control. First recall the following definition:

Definition 1.2. A set $E \subset \mathbb{R}^n$ is *porous at a point $x \in E$* if there is a $c > 0$ and there is a sequence $y_n \to 0$ such that the balls $B(x+y_n, c|y_n|)$ are disjoint from E. The set E is *porous* if it is porous at each of its points, and it is called *σ-porous* if it is a countable union of porous sets.

Theorem 1.3 ([3]). *Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a Lipschitz function. Then the set of those points at which f is not differentiable but it is differentiable in n linearly independent directions is σ-porous.*

It follows from Lebesgue's density theorem that σ-porous sets have Lebesgue measure zero. Therefore as an immediate corollary we obtain:

Theorem 1.4 (Rademacher). *Every Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is differentiable almost everywhere.*

The converse direction, i.e. the description of those sets $E \subset \mathbb{R}^n$ for which there is a non-differentiable Lipschitz function, is much harder. D. Preiss proved that the converse of Rademacher's theorem is false, already in dimension 2:

Theorem 1.5 ([9]). *There is a Lebesgue null set $E \subset \mathbb{R}^2$ such that every Lipschitz function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is differentiable in at least one point of E.*

Unlike in the classical Lebesgue and Rademacher theorem, Preiss's result is not an 'almost everywhere' result, he does not show that the function is differentiable at 'most' of the points $x \in E$. Indeed this is not possible. We prove the following theorem:

Theorem 1.6. *For every Lebesgue null set $E \subset \mathbb{R}^2$ there is a Lipschitz function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ which is not differentiable at any point $x \in E$.*

This theorem says that for every Lebesgue null set there are two real-valued Lipschitz functions, namely, the coordinate functions of f, such that at each $x \in E$ at least one of the two functions are non-differentiable.

Remark. In [9] the result is proved not only in $\mathbb{R}^2$, but in every Banach space with a smooth norm. Preiss's set E is dense. In a recent paper [5], M. Doré and O. Maleva constructed a closed (and hence nowhere dense) null set with the same property: in every Banach space X with separable dual there exists a closed bounded set of Hausdorff dimension one containing a Fréchet-differentiability point of every Lipschitz function $f\colon X \to \mathbb{R}$.

Let $E \subset \mathbb{R}^n$. It is immediate from the definition that a set E is porous at $x \in E$ if and only if the Lipschitz function $f(x) = \operatorname{dist}(x, E)$ is non-differentiable at x. Of course σ-porous sets cannot fully describe non-differentiability sets of Lipschitz functions (not even in $\mathbb{R}$, since not all Lebesgue null sets of $\mathbb{R}$ are σ-porous). But by Theorem 1.3, in order to find all Lebesgue null sets for which there is a non-differentiable Lipschitz function, it is enough to consider functions not having enough many directional derivatives.

From the point of view of differentiability problems, the sets that are the most negligible are the sets of points at which a Lipschitz function may be differentiable in no direction. We show that these sets form a σ-ideal. We call them *uniformly purely unrectifiable*. Notice that uniformly purely unrectifiable sets are purely unrectifiable, i.e. they are null on every rectifiable curve, since a Lipschitz function is differentiable in the tangent direction at a.e. point of a curve. We will see later that uniformly purely unrectifiable sets have the (possibly only formally) stronger property that they can be covered by an open set which is small on many curves simultaneously.

For simplicity, consider just Lipschitz functions $f\colon \mathbb{R}^2 \to \mathbb{R}^m$. We will show that if f is not differentiable at the points of $E \subset \mathbb{R}^2$, then at each point $x \in E$ except for a uniformly purely unrectifiable set, there is a *unique* differentiability direction $\tau(x)$ of f. Moreover, this direction is determined by the geometry of the set E, it is independent of the function f: for any other Lipschitz function g, the direction constructed using f and g agree at each point of E except for a uniformly purely unrectifiable set. Indeed, if E is contained in the non-differentiability set of both $f\colon \mathbb{R}^2 \to \mathbb{R}^{m_1}$ and $g\colon \mathbb{R}^2 \to \mathbb{R}^{m_2}$, then the direction τ defined by the function $h = (f, g)\colon \mathbb{R}^2 \to \mathbb{R}^{m_1+m_2}$ must coincide with the directions defined by f or g, whenever f, g and h have a unique direction of differentiability.

Using also Theorem 1.6, we obtain:

Corollary 1.7. *For every planar Lebesgue null set E, at each point $x \in E$ there is a direction $\tau(x)$ with the following property: every Lipschitz function $f\colon \mathbb{R}^2 \to \mathbb{R}^m$ is differentiable in the direction $\tau(x)$ at every $x \in E$, except at a uniformly purely unrectifiable set of points. This direction is determined uniquely, except for a uniformly purely unrectifiable set.*

Remark. There are null sets which are very far from being purely unrectifiable. For instance, R. O. Davies showed in [4] that every Borel set $B \subset \mathbb{R}^2$ can be covered by infinite straight lines without increasing its Lebesgue measure. One can even put continuum many lines through each of the points of B so that the

union of these lines has the same measure as B. Now if $B = B_0$ is, say, a point, applying Davies's theorem iteratively, we can find $B_0 \subset B_1 \subset B_2 \subset \dots$ such that each B_k has continuum many lines through the points of B_{k-1}, and the sets B_k are Lebesgue null. Then $\bigcup B_k$ is also Lebesgue null, and it has continuum many lines through each of its points. What could be τ on $\bigcup B_k$? Since Lipschitz functions are differentiable along lines, at each line of the construction, τ must agree with the direction of the line at a.e. of its points. But there are continuum many lines at each point, how can we choose only one of these, so that along any given line at a.e. point we choose the direction of the given line and not one of the others?

Now, consider Lipschitz functions on $\mathbb{R}^n$.

Notation. We denote by $\mathcal{N}_{n,k}$ the σ-ideal of subsets of $\mathbb{R}^n$ generated by sets for which there is a Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}$ differentiable in at most k linearly independent directions.

So $\mathcal{N}_{n,0}$ are exactly the uniformly purely unrectifiable sets, while $\mathcal{N}_{n,n-1}$ are the non-differentiability sets we are mainly interested in.

Since a Lipschitz function is differentiable in the tangent directions of any k-rectifiable set at $\mathcal{H}^k$-almost all of its points, therefore $\mathcal{N}_{n,k-1}$ sets are k-purely unrectifiable, i.e. they meet every k-rectifiable set in an $\mathcal{H}^k$-null set.

As a refinement of the above observations on directions of differentiability in the plane, we will show that whenever $E \in \mathcal{N}_{n,k}$, there is $\tau\colon E \to G(n,k)$ such that for all $x \in E$ except those belonging to an $\mathcal{N}_{n,k-1}$ set, every Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is differentiable in the direction $\tau(x)$.

Definition 1.8. $\tau\colon E \to G(n,k)$ is called a *k-dimensional tangent field* of a set E if every Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is differentiable in the direction $\tau(x)$ at all $x \in E$ except those belonging to an $\mathcal{N}_{n,k-1}$ set.

Theorem 1.9. *Every set $E \in \mathcal{N}_{n,k}$ has a k-dimensional tangent field. Moreover, the tangent field is unique up to an $\mathcal{N}_{n,k-1}$ set.*

It is easy to see that:

Proposition 1.10. *The set of (directional) non-differentiability of a Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}$ can be written as a countable union of sets E, for each of which we may find a direction u and numbers $a < b$ such that*

$$\liminf_{t\to 0} \frac{f(x+tu)-f(x)}{t} < a < b < \limsup_{t\to 0} \frac{f(x+tu)-f(x)}{t}.$$

Since our f is Lipschitz, such set is null not only on every line in direction u, but also on every curve $\gamma\colon \mathbb{R} \to \mathbb{R}^n$ provided that $|\gamma' - u|$ is small enough.

We can do slightly better: if $\delta > 0$ is small enough, for every $\varepsilon > 0$ there is an open set $G \supset E$ such that the length of $G \cap \gamma$ is less than ε for every curve $\gamma\colon \mathbb{R} \to \mathbb{R}^n$ with $|\gamma' - u| < \delta$.

This observation motivates the following definition. Given a convex cone C, we may define the C-width of an open set G as the supremum of the lengths of $\gamma \cap G$ where the supremum is taken over all Lipschitz curves γ that 'go in the direction of C', i.e. for which $\gamma'(t) \in C$ for a.e. t. Then we define the C-width for general sets as the infimum of the C-widths of open sets containing it. In fact, our definition of the width is slightly more complicated: instead of the length we use a technically more convenient measure (that also depends on a vector $e \in \operatorname{int}(C)$) of the part of the curve that lies in the set G. (See later, Definition 1.14.)

Using this notion of width, an equivalent description of the tangent field of a set can be obtained without referring to non-differentiability sets and non-differentiability directions of Lipschitz functions:

Definition 1.11. If $E \subset \mathbb{R}^n$, we say that the mapping $\tau\colon E \to G(n,k)$ is a *k-dimensional tangent field* of E if for every cone C, the set of those points $x \in E$ for which $\tau(x) \cap C = \{0\}$ has C-width zero.

This defines the same tangent field as Definition 1.8: one can show that the family of those subsets of $\mathbb{R}^n$ that admit a k-dimensional tangent field according to Definition 1.11 coincides with the σ-ideal $\mathcal{N}_{n,k}$, and also that the two tangent fields coincide.

According to Proposition 1.10 (and paragraphs preceding it), the set where f is not differentiable can be covered by countably many sets, each of which has width zero with respect to some cone.

We do not know whether this is a full description, i.e. we do not know whether the non-differentiability sets of Lipschitz functions (i.e. those sets that admit an $(n-1)$-tangent field) are exactly described by the property that they can be covered by countably many sets, each of which has width zero with respect to some cone. It is not very hard to show, using Definition 1.11, that the existence of an $(n-1)$-tangent field of a set is equivalent to the property that for every $\varepsilon > 0$ the set can be covered by a finite number of sets each of which has width zero with respect to some cone that is only ε-far from a halfspace.

Our results include:

Theorem 1.12. *For every set $E \subset \mathbb{R}^n$, the following are equivalent:*

(i) There is a Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^n$ that is non-differentiable at any point of E.

(ii) There is a sequence (possibly infinite) of Lipschitz functions $f_j\colon \mathbb{R}^n \to \mathbb{R}$ such that at every point of E at least one of the f_j is non-differentiable.

(iii) The set E is in $\mathcal{N}_{n,n-1}$.

(iv) The set E has an $(n-1)$-tangent field.

(v) If $n \leq 2$: E has Lebesgue measure zero.

We do not know whether every Lebesgue null set is in $\mathcal{N}_{n,n-1}$ for $n > 2$. And we do not know whether it is true that for every $m < n$ there is a null set $E \in \mathbb{R}^n$ such that every Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at some point of E. Preiss proved in [9] that the answer is 'yes' for $1 = m < n$. Doré and Maleva in [6], building heavily on methods due to Lindenstrauss, Preiss and Tišer in [8] in their study of differentiability problems in infinite dimensional Banach spaces, proved that the answer is also yes for $2 = m < n$. But their current methods do not work for $m \geq 3$.

So far we didn't say anything about how we can construct a non-differentiable function for a given (small) set E. This is much harder than the other direction, i.e. showing that the set of points of non-differentiability must be small. In dimension 1 it is easy, and one may try to use the 1-dimensional proof as a guidance. One could even consider generalising the more precise description of the sets of non-differentiability of Lipschitz functions $f\colon \mathbb{R} \to \mathbb{R}$ due (with slightly worse constants) to Zahorski [11]. (See [7] for a more recent proof.)

Theorem 1.13 (Zahorski). *For any G_δ set $E \subset \mathbb{R}$ of Lebesgue measure zero there is a Lipschitz function $f\colon \mathbb{R} \to \mathbb{R}$ with $\operatorname{Lip}(f) \leq 1$ which is differentiable at every point $x \notin E$ and*

$$\liminf_{t\to 0} \frac{f(x+t)-f(x)}{t} = -1 < 1 = \limsup_{t\to 0} \frac{f(x+t)-f(x)}{t}$$

for every $x \in E$.

Recall that a set is G_δ if it is an intersection of countably many open sets. Recall also that, by adding together suitable multiples of functions obtained by this theorem for G_δ sets E_i, Zahorski showed that $E \subset \mathbb{R}$ is the set of points of non-differentiability of some Lipschitz function $f\colon \mathbb{R} \to \mathbb{R}$ if and only if E is of Lebesgue measure zero and of type $G_{\delta\sigma}$ (a union of countably many G_δ sets).

So let us see how one can construct a Lipschitz function $f\colon \mathbb{R} \to \mathbb{R}$ non-differentiable at the points of the given Lebesgue null set E. We recursively find open sets $G_1 \supset G_2 \supset \cdots \supset E$ so small that G_k is small in every component of G_{k-1}. (For example, $|G_k \cap C| < 2^{-k}|b-a|$ for any component $C = (a,b)$ of G_{k-1}.) Let $f_k(x)$ denote the measure of $(-\infty, x) \cap G_k$. Then $f_k'(x) = 1$ at each point $x \in G_k$, but the slope $(f_k(b) - f_k(a)/(b-a)$ is close to 0. Using this it is easy to check that $f(x) = \sum(-1)^k f_k(x)$ is not differentiable at any point of $\bigcap G_k$. If E is G_δ and $\varepsilon > 0$, it is not difficult to choose the G_k so that, defining $f(x) = \sum \lambda_k f_k(x)$ where $|\lambda_k| < \varepsilon$ and the partial sums of the λ_k oscillate between ± 1, we get a function that almost satisfies the statement of Theorem 1.13. However, at the points of $\mathbb{R} \setminus E$ we would only get that the upper and lower derivatives of f differ by no more than 2ε, not that f is differentiable. We are in fact able to find a higher dimensional analogue of this construction. Recall however that Theorem 1.13 is proved in a different way, and that the weaker statement that we have just indicated is not sufficient for showing the full description of non-differentiability sets mentioned above.

As a higher dimensional analogue of the functions f_k, for an open set $G \subset \mathbb{R}^n$ of (small) C-width w and unit vector e from the interior of C, we construct a function $\omega\colon \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{Lip}(\omega)$ is bounded by a constant depending on C and e, $\omega(y) \geq \omega(x)$ if $y - x \in C$, $\omega(x+te) = \omega(x)+t$ if the segment $[x, x+te]$ lies in G, and $0 \leq \omega(x) \leq w$ for all $x \in \mathbb{R}^n$.

The function ω can be used to construct non-differentiable functions, in a similar way as the functions f_k were used in dimension 1. Indeed, ω has directional derivative 1 in the direction e at each $x \in G$, but from the more global point of view ω looks like having derivative zero.

The technical details of the construction are quite complicated. They may be somewhat simplified in the case of sets $E \in \mathcal{N}_{n,0}$. Given any vector e, we choose an open set $G \supset E$ with small C-width where C is close to the halfspace $\{x : \langle x, e\rangle \geq 0\}$. The function $\langle x, e\rangle - \omega(x)$ sees, from every point of G, some points in the direction e with slope almost one, but has local Lipschitz constant close to zero on G. This allows us to iterate the construction locally. Moving also the vectors e through a dense subset of the unit sphere, we get a function which is non-differentiable at any point of E in any direction. More precisely, here is our definition and the results we prove:

Definition 1.14. Let C be a convex cone and let e be a unit vector in C.

(i) We define $M = M_{C,e}\colon \mathbb{R}^n \to \mathbb{R}$ by

$$M(x) = \sup\{\lambda \in \mathbb{R} : x - \lambda e \in C\}.$$

(ii) The *C-width* $w(G) = w_{C,e}(G)$ of an open set $G \subset \mathbb{R}^n$ is defined as the supremum of the numbers

$$\int_{\{t:\gamma(t)\in G\}} M(\gamma'(t))\, dt$$

among all Lipschitz curves $\gamma\colon \mathbb{R} \to \mathbb{R}^n$ which go in the direction of C.

(iii) For a general set $E \subset \mathbb{R}^n$, $w(E)$ is the infimum of $w(G)$ among all open sets G which contain E.

(iv) Let $G \subset \mathbb{R}^n$ be an open set of finite width. For every point $x \in \mathbb{R}^n$ we set $\omega(x) = \omega_{G,C,e}(x)$ as the supremum of the numbers

$$-\lambda + \int_{t\in[a,b],\gamma(t)\in G} M(\gamma'(t))\, dt$$

among all $a, b \in \mathbb{R}$, $\lambda \geq 0$ and $\gamma : [a, b] \to \mathbb{R}^n$ such that $\gamma(b) - x = \lambda e$ and γ goes in the direction of C.

We use this function ω to prove:

Theorem 1.15. *For every $\tilde{\varepsilon} > 0$ and for every set E which is G_δ and uniformly purely unrectifiable there is a function $f\colon \mathbb{R}^n \to \mathbb{R}$ such that*

(i) $\operatorname{Lip}(f) = 1$;

(ii) *f is $\tilde{\varepsilon}$-differentiable on $\mathbb{R}^n \setminus E$, that is, for every $x \in \mathbb{R}^n \setminus E$ there is $r > 0$ and a vector u such that*

$$|f(x) - f(y) - \langle u, y - x\rangle| \le \tilde{\varepsilon}|y - x| \quad \text{for all} \quad y \in B(x, r),$$

(iii) *for every $x \in E$, $\eta \in B(0,1) \subset \mathbb{R}^n$ and $\varepsilon > 0$ there is an $r < \varepsilon$ such that*

$$|f(y) - f(x) - \langle \eta, y - x\rangle| \le \varepsilon r \quad \text{for all} \quad y \in B(x, r).$$

In particular, f is not differentiable at the points of E, it is not even ε-differentiable for any $\varepsilon < 1$.

Since every uniformly purely unrectifiable set is contained in a G_δ uniformly purely unrectifiable set, this indeed shows that for every $\mathcal{N}_{n,0}$ set there is a Lipschitz function that is non-differentiable in any direction. However this result does not provide Zahorski-type exact description of sets of non-differentiability in any direction (which, by analogy, one would conjecture to be $\mathcal{N}_{n,0}$ sets of type $G_{\delta\sigma}$), since we do not know (in dimension $n > 1$) whether (ii) of Theorem 1.15 can be replaced by the condition that f is differentiable on $\mathbb{R}^n \setminus E$.

By a rather delicate induction with respect to k (which is where we need the condition (ii) of Theorem 1.15) we show that the sets of points of k-dimensional differentiability can be characterised as follows:

Theorem 1.16. *(i) Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a Lipschitz function, and for each $x \in \mathbb{R}^n$ choose $\tau(x)$ to be a maximal dimensional subspace such that the restriction of f to $x+\tau(x)$ is differentiable at x. For each $0 \le k \le n-1$, let E_k denote the set of those points at which $\dim \tau(x) = k$. Then $E_k \in \mathcal{N}_{n,k}$.*

(ii) Let $E_k \subset \mathbb{R}^n$ be an $\mathcal{N}_{n,k}$ set for some $0 \le k \le n-1$. Then there is a Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^{k+1}$ and a k-tangent field τ of E_k such that f is not differentiable at any $x \in E_k$ in any direction e that is orthogonal to $\tau(x)$.

We can make (ii) of Theorem 1.16 more quantitative. Again, this is a weaker analogy of Theorem 1.13, which is needed for induction and to which the same remarks as to the case $k = 0$ apply.

Theorem 1.17. *For each $0 \le k < n$ there is a constant $c_{n,k} > 0$ such that, whenever $l > k$, $\varepsilon > 0$ and E is a G_δ, $\mathcal{N}_{n,k}$ subset of $\mathbb{R}^n$, then there is a function $f\colon \mathbb{R}^n \to \mathbb{R}^l$ with $\operatorname{Lip}(f) \le 1$ which is ε-directionally differentiable at*

every point of $\mathbb{R}^n \setminus E$ *and has the property that for every* $x \in E$ *there are* k*-dimensional linear subspaces* V, W *of* $\mathbb{R}^n, \mathbb{R}^l$*, respectively, so that for any unit vectors* $v \in V^\perp$ *and* $w \in W^\perp$*,*

$$\limsup_{t \searrow 0} \frac{\langle f(x+tv) - f(x), w\rangle}{t} - \liminf_{t \searrow 0} \frac{\langle f(x+tv) - f(x), w\rangle}{t} \geq c_{n,k}.$$

According to (iii) of Theorem 1.15, $c_{n,0} = 2$. We do not know whether $c_{n,k} = 2$ for $k > 0$.

We finish this section by showing that for differentiability with respect to a measure (instead of at every point of a given set) it is sufficient to consider real-valued functions:

Theorem 1.18. *Let* μ *be a* σ*-finite Borel measure on* $\mathbb{R}^n$*.*

(i) Every real-valued Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}$ *is differentiable* μ*-almost everywhere, if and only if every set in* $\mathcal{N}_{n,n-1}$ *is* μ*-null.*

(ii) On the other hand, if an $\mathcal{N}_{n,n-1}$*-set has positive* μ*-measure, then there is a Lipschitz function* $f\colon \mathbb{R}^n \to \mathbb{R}$ *which is non-differentiable* μ*-almost everywhere on this set.*

In particular, for every singular probability measure μ *in the plane there is a Lipschitz function* $f\colon \mathbb{R}^2 \to \mathbb{R}$ *which is non-differentiable* μ*-almost everywhere.*

This nicely complements the result of Preiss mentioned before, according to which there is a null set $E \subset \mathbb{R}^2$ such that every Lipschitz function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is differentiable in at least one point of E.

As we have already pointed out, the proof of Theorem 1.16 is rather involved. However, Theorem 1.18 may be proved in a simpler way, closer to the argument that we indicated for Theorem 1.15. Recall that the key point of this argument was that for an open set G of small C-width and $e \in C$ we constructed a function ω with directional derivative 1 in the direction e at each $x \in G$, but looking like having derivative zero from the global point of view. To prove Theorem 1.15, we needed only one such G (as it contained the whole set E) while to prove Theorem 1.16 we need several of them which may overlap and so constructions that we need to do cannot be independent. However, to show Theorem 1.18, we may throw away sets of small measure, and so achieve that the sets G in which we have to construct the function ω are in positive distance from each other. These constructions may still be handled independently, resulting in a reasonably accessible proof.

2. Structure of Null Sets and Other Problems

In this section we list various results that can be proved using similar techniques and ideas as the ones we use for the characterisation of non-differentiability of Lipschitz functions.

2.1. Tangent of null sets. In the planar case, we know that the σ-ideal $\mathcal{N}_{2,1}$ and the σ-ideal of Lebesgue null sets coincide, i.e. every planar Lebesgue null set admits a 1-tangent field. We do not know if the same is true in higher dimension. However, there is another, weaker notion of tangent fields that can be defined for any Lebesgue null set in $\mathbb{R}^n$:

Definition 2.1. Given a set $E \subset \mathbb{R}^n$, we say that a Borel measurable map $\tau\colon E \to G(n,k)$ defines a *weak k-tangent field* to E if for every k-rectifiable set S, $\operatorname{Tan}(S,x) = \tau(x)$ for $\mathcal{H}^k$-a.e. $x \in S \cap E$.

Notice that in this definition we had to include a measurability assumption. It was not needed in Definition 1.8 since the tangent field defined there is automatically Borel measurable (after a modification on an $\mathcal{N}_{n,k-1}$ set). However, under the continuum hypothesis one can define a non-measurable weak k-tangent field by ordering k dimensional C^1 surfaces in $\mathbb{R}^n$ into S_α, $\alpha < \omega_1$ and defining $\tau(x)$ as the tangent space of S_α at x where α is the first ordinal for which $x \in S_\alpha$.

It follows from the definition that, given a set $E \subset \mathbb{R}^n$, the weak k-tangent field, provided that it exists, is uniquely defined up to k-purely unrectifiable subsets of E (recall that the k-tangent field is uniquely defined up to an $\mathcal{N}_{n,k-1}$ set). Also, if a set admits a k-tangent field then it is also a weak k-tangent field. We do not know (even in the planar case for $k=1$) whether the σ-ideal $\mathcal{N}_{n,k-1}$ coincides with the σ-ideal generated by G_δ (or Borel, or analytic) k-purely unrectifiable sets, and we do not know in dimensions $n > 2$ whether every set admitting a weak k-tangent field also admits a k-tangent field. However, we can prove that:

Theorem 2.2. *Any set $E \subset \mathbb{R}^n$ of Lebesgue measure zero admits a weak $(n-1)$-tangent field.*

This result can be understood as saying the rather mysterious fact that one can prescribe in which direction an $(n-1)$-surface meets a null set E, without knowing the surface itself. The mystery would deepen if, for example, one had a purely 1-unrectifiable set in $\mathcal{N}_{n,1} \setminus \mathcal{N}_{n,0}$: this set would have uniquely prescribed directions that would not be possible to describe by meeting with curves.

2.2. Covering by Lipschitz slabs and intersecting by curves. The notion 'C-width' can be defined in the following, equivalent way. Given a cone C and a vector $e \in \operatorname{int}(C)$, if E is a 'C-Lipschitz set', i.e. $E \cap (x+C) = \{0\}$ and E meets each line of direction e in exactly one point, then we call the set between E and its shifted copy $E + we$ ($w > 0$) a *C-Lipschitz slab of width w*. If $K \subset \mathbb{R}^n$ is compact, we may define its *C-width* as the infimum of the total width of families of C-Lipschitz slabs covering it. If $G \subset \mathbb{R}^n$ is open, then we define its C-width as the supremum of C-widths of compact sets contained in it, and finally if $E \subset \mathbb{R}^n$ is arbitrary, then its C-width is defined as the infimum of the C-widths of open sets containing it.

In our original definition of C-width, we measured the part of the curve γ that lies in the set G (i.e. we chose the function $M(x)$ in (i) of Definition 1.14) in such a way that we obtain *exactly* the same width as the one defined using C-Lipschitz slabs.

So a compact set has C-width zero if it can be covered by C-Lipschitz slabs of arbitrary small total width. In particular, in $\mathbb{R}^2$, every compact Lebesgue null set is in $\mathcal{N}_{2,1}$, therefore it can be covered by Lipschitz slabs of arbitrary small total width. In fact, in the plane one can cover any null set, and it is enough to use the coordinate directions and Lipschitz graphs with Lipschitz constant one. We show the following:

Theorem 2.3. *Every set $E \subset [0,1]^2$ of measure $0 \le m < ab$ is the union of two sets $E = A \cup B$, where A has C-width less than a for $C = \{(x,y) : |x| > |y|\}$ and B has C-width less than b for $C = \{(x,y) : |y| > |x|\}$.*

That is, there are Lipschitz functions $f_i \colon \mathbb{R} \to \mathbb{R}$, $g_j \colon \mathbb{R} \to \mathbb{R}$ with Lipschitz constant 1 and $w_i, w_j > 0$ with $\sum_i w_i < a$, $\sum_j w_j < b$, such that

$$A \subset \bigcup_i \{(x,y) : f_i(x) \le y \le f_i(x) + w_i\} \quad B \subset \bigcup_j \{(x,y) : g_j(y) \le x \le g_j(y) + w_j\}.$$

This can be used e.g. to show that there is a 1-Lipschitz function $f \colon \mathbb{R} \to \mathbb{R}$ whose graph $(x, f(x))$ or $(f(x), x)$ meets E in length at least $m^{1/2}$. The analogous result is also true in higher dimension:

Theorem 2.4. *For every set $E \subset [0,1]^n$ of measure m there is a Lipschitz curve (with a fixed Lipschitz constant that depends only on the dimension n) that meets E in length at least $c_n m^{1/n}$.*

Here a curve means the graph of a map from one of the coordinate axis into its orthogonal complement. We do not know whether there is a k-dimensional Lipschitz surface (where surfaces are understood similarly) that meets E in $\mathcal{H}^k$-measure $c_{k,n} m^{k/n}$.

2.3. Mappings onto balls and weak derivatives.

Among the problems exploring the geometric structure of sets with positive Lebesgue measure, the following one, proposed by M. Laczkovich, is particularly interesting:

Problem 2.5. *Given a set $E \subset \mathbb{R}^n$ of positive Lebesgue measure, is there a Lipschitz function $f \colon \mathbb{R}^n \to \mathbb{R}^n$ which maps E onto a set with non-empty interior (or, equivalently, that maps E onto a ball)?*

Without loss of generality we can assume that E is compact. In dimension $n = 1$, $f(x) = |(-\infty, x) \cap E|$ maps E onto an interval and $\mathbb{R} \setminus E$ onto a countable set.

P. Jones called our attention to a result of N. X. Uy in [10]:

Theorem 2.6 ([10]). *For every compact set $E \subset \mathbb{R}^2$ of positive Lebesgue measure there is a non-constant complex-valued Lipschitz function that is holomorphic everywhere outside E (including infinity).*

If we identify $\mathbb{C}$ and $\mathbb{R}^2$, we obtain a mapping $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ that is orientation preserving and open on the complement of E; using degree theory it follows that $f(E) \supset f(\mathbb{R}^2 \setminus E) \supset$ a ball. This gives a positive answer to Problem 2.5 in dimension $n = 2$.

In dimension $n = 2$ a completely different construction can also be obtained using our function ω from (iv) Definition 1.14 (more precisely, the function $u(x) = x - \omega(x)e$, whose distance from the identity is small). Instead of constructing an open mapping on $\mathbb{R}^2 \setminus E$, we show that close to a density point of E a Lipschitz perturbation of the identity can be found which maps $\mathbb{R}^2 \setminus E$ onto a 1-rectifiable set (and consequently, it maps E onto a set of non-empty interior):

Theorem 2.7. *For $n = 1, 2$ and for every $E \subset \mathbb{R}^n$ of positive Lebesgue measure there is an orientation-preserving Lipschitz mapping $f\colon \mathbb{R}^n \to \mathbb{R}^n$ such that $f(E) = [0,1]^n$ and $f(\mathbb{R}^2 \setminus E)$ is $(n-1)$-rectifiable.*

Unfortunately none of these methods are powerful enough to construct such a mapping in higher dimension; the question in dimensions $n \geq 3$ remains open. It may be true that, in any dimension, there is a Lipschitz perturbation of the identity that maps $\mathbb{R}^n \setminus E$ onto an $(n-1)$-rectifiable set.

Another use of ω is the following. Let μ be a measure such that $\mu(S) > 0$ for some $S \in \mathcal{N}_{n,n-1}$, and let E be a subset of S with $\mu(E) > 0$ of C-width zero for some cone C. Let ω_j denote the function ω for $w = 1/j$. Then the functions $\omega_j\colon \mathbb{R}^n \to \mathbb{R}$ have uniformly bounded Lipschitz constants, they converge to constant 0 as j tends to infinity, and $\omega_j'(x;e) \geq 0$ everywhere and $\omega_j'(x;e) = 1$ for $x \in E$.

A moment's reflection shows that ω_j' cannot converge to $0 = \omega'$ in any weak sense with respect to μ (and a straightforward smoothing argument can make them C^1). Therefore, for a measure μ in $\mathbb{R}^n$, weak derivatives of Lipschitz functions may be defined iff μ is absolutely continuous with respect to $\mathcal{N}_{n,n-1}$, hence iff every Lipschitz function is differentiable μ-almost everywhere. For $n = 2$ we know that the above holds iff μ is absolutely continuous with respect to the Lebesgue measure. This answers a problem due to G. Mokobodzki.

2.4. Tangents of measures.

Alberti proved in [1] the so-called 'rank-one property' of BV functions:

Theorem 2.8 ([1]). *Let u and v be BV functions on $\mathbb{R}^n$. Then the direction of the gradients of u, v agree μ-a.e. whenever the measure μ is singular, and absolutely continuous with respect to the variation of the gradients of both u and v.*

This result can be understood as saying that certain class of $\mathbb{R}^n$-valued measures in $\mathbb{R}^n$, namely those that arise as singular parts of derivatives of BV functions, have a.e. uniquely defined normal directions and so also 'tangent'

hyperplanes. The question naturally arises: for what measures is our $(n-1)$-dimensional tangent field uniquely defined almost everywhere? Is it the same as the hyperplane defined via derivatives of BV functions?

The measure has to be concentrated on $\mathcal{N}_{n,n-1}$ and it has to be absolutely continuous with respect to $\mathcal{N}_{n,n-2}$. Since sets from $\mathcal{N}_{n,n-2}$ are purely $(n-1)$-unrectifiable, for the later requirement it suffices that the measure is absolutely continuous with respect to purely $(n-1)$-unrectifiable sets. The former requirement would be equivalent to singularity if $\mathcal{N}_{n,n-1}$ coincided with Lebesgue null sets, which we do not know. But the methods used to prove it when $n = 2$ are powerful enough to show that every Lebesgue null set in $\mathbb{R}^n$ is a union of a set from $\mathcal{N}_{n,n-1}$ and a purely $(n-1)$-unrectifiable set. So it suffices to assume that the measure is singular, and absolutely continuous with respect to purely $(n-1)$-unrectifiable sets.

Definition 2.9. A measure on $\mathbb{R}^n$ is called *k-rectifiable* if it is absolutely continuous with respect to $\mathcal{H}^k|_E$, where $E \subset \mathbb{R}^n$ is a k-rectifiable set. Measures which can be represented as integral combinations $\mu = \int \mu_t \, dP(t)$ of k-rectifiable measures μ_t are called *k-rectifiably representable.*

Theorem 2.10. *A measure μ is k-rectifiably representable if and only if $\mu(E) = 0$ for every k-purely unrectifiable set E.*

Definition 2.11. Given a k-rectifiably representable measure μ on $\mathbb{R}^n$, $\tau\colon \mathbb{R}^n \to G(n,k)$ defines a *k-tangent field of μ*, if for every representation $\mu = \int \mu_t \, dP(t)$ where μ_t is supported on a k-rectifiable set E_t, there holds $\mathrm{Tan}(E_t, x) = \tau(x)$ for μ_t-a.e. x and P-a.e. t.

The k-tangent field, if it exists, is uniquely determined up to μ-negligible sets. We show that:

Theorem 2.12. *An $(n-1)$-rectifiably representable measure admits an $(n-1)$-tangent field if and only if it is singular.*

Singular parts of derivatives of BV functions are $(n-1)$-rectifiably representable, and indeed, the hyperplane orthogonal to the gradient and the $(n-1)$-tangent field of these measures coincide.

We finish this section by saying that, applying a version of Radon-Nýkodim Theorem, we show that

Theorem 2.13. *Every measure μ on $\mathbb{R}^n$ can be uniquely decomposed as*

$$\mu = \mu_n + \mu_{n-1} + \cdots + \mu_0,$$

where each μ_k is a k-rectifiably representable measure supported on a $(k+1)$-purely unrectifiable set.

We do not know whether the measure μ_k admits a k-tangent field for $k < n-1$.

3. Combinatorial Connections

Combinatorial connections of our results were first noted by Matoušek. He observed that a part of our proof of Laczkovich's problem is similar to the proof of the Erdős-Szekeres Theorem, and he recognised that this part may be replaced by its corollary: for any planar set M having m^2 points there is a function $\psi\colon \mathbb{R} \to \mathbb{R}$ with $\operatorname{Lip}(\psi) \leq 1$ such that one of the sets

$$\{(x,y) \in M : y = \psi(x)\} \quad \text{or} \quad \{(x,y) \in M : x = \psi(y)\}$$

has at least m points.

Some of the results on which our proofs are based exploited this connection and may be considered as a continuous analogy of the Erdős-Szekeres or Dilworth Theorems. For example, if a set E admits a k-tangent field then for every decomposition $G(n,k) = \bigcup A_j$ there corresponds a partition $E \subset \bigcup E_j$, where E_j contains those $x \in E$ for which $\tau(x) \in A_j$. By definition, the set E_j has width 0 with respect to any proper closed convex cone C for which $C \cap S = \{0\}$ for all $S \in A_j$. That is, we can decompose E into parts that can be covered by Lipschitz slabs of arbitrary small total width. Discrete analogue of this statement says that, in the plane, a finite set of points can always be covered by a small number of Lipschitz curves of given directions (and then of course one of them must contain many points).

The relation to the combinatorial results becomes even more apparent if we consider weak k-tangent fields. Look at only the special case $k = n-1$, and suppose that $E \subset [0,1]^n$ is Lebesgue null and compact. Then we can approximate E by a grid: it intersects $o(N^n)$ out of N^n subcubes of $[0,1]^n$. Let C be a convex cone, and consider the partial order on $\mathbb{R}^n$ defined by $x_1 \prec x_2 \iff x_2 - x_1 \in C$. By Dilworth Theorem, the set of the centres of the cubes intersecting E can be covered by $o(N^{n-1})$ chains and $o(N)$ antichains. Chains are curves going in the direction of C and antichains are C-Lipschitz surfaces. Since E lies in a $O(1/N)$ neighbourhood of the set of the centres of the cubes, it is covered by $o(N^{n-1})$ 'tubes' going in the direction C and by $o(N)$ Lipschitz slabs of width $O(1/N)$, i.e. by tubes of arbitrary small total cross-sectional volume and by slabs of arbitrary small total width. The set covered by tubes meets C-Lipschitz surfaces in a set of small $\mathcal{H}^{n-1}$ measure, and the set covered by slabs meets curves going in the direction C in small length.

This decomposition leads to a weak $(n-1)$-tangent field as we let the angle of C tend to a halfspace and its direction run through a dense set of directions. For other results we would need to cover by slabs only, and we do not know if this is always possible, except for the 2-dimensional case where there is no difference between tubes and slabs.

Many results presented here are connected to a possibility of decomposing certain small sets, or perhaps even all Lebesgue null sets, in a way reminiscent of the decompositions of finite sets in combinatorial results. As we have seen above, the existence of a weak $(n-1)$-tangent field is a direct corollary of Dilworth

Theorem. For other problems we need a much finer, continuous version of the combinatorial results whose proofs also use techniques that are not available in the discrete world, they are purely analytic.

There are also problems that could be solved using discrete decomposition results, but we do not know if the discrete versions are true. Matoušek conjectured a higher dimensional variant of the Erdős-Szekeres Theorem that would fully solve Laczkovich's problem. This conjecture was disproved by Tardos. One can however modify his conjecture so that it would imply a positive answer to our main problem (all Lebesgue null sets would belong to $\mathcal{N}_{n,n-1}$):

Conjecture 3.1. For any set $M \subset \mathbb{R}^n$ having m^n points there is a function $\psi\colon \mathbb{R}^{n-1} \to \mathbb{R}$ with $\mathrm{Lip}(\psi) \le C_n$ and *an orthonormal system of coordinates* such that the set

$$\{(x_1, \dots, x_n) \in M : x_n = \psi(x_1, \dots, x_{n-1})\}$$

has at least $c_n m^{n-1}$ points.

This problem is open. We only show that, unlike in the plane, the coordinate systems cannot be restricted to permutations of the standard coordinate system, not even in $\mathbb{R}^3$:

Theorem 3.2. *For every Lipschitz constant L and for every $\varepsilon > 0$ there exists a finite set $M \subset \mathbb{R}^3$ of m^3 points, such that for every $\phi\colon \mathbb{R}^2 \to \mathbb{R}$ with $\mathrm{Lip}(\phi) < L$, in the standard coordinate system in $\mathbb{R}^3$, all the three graphs $x = \phi(y,z)$, $y = \phi(x,z)$ and $z = \phi(x,y)$ contain less than εm^2 points of M.*

However, the *dyadic* analogue of Conjecture 3.1 is true in any dimension, even in the standard coordinate-system for Lipschitz mappings with constant 1.

Consider the unit cube $Q = [0,1]^n \subset \mathbb{R}^n$. A cube in Q is called a *dyadic cube* of size $1/2^k$, if it is obtained by dividing Q to 2^{kn} subcubes of equal sizes in the obvious manner. Let Q_0 be the set of points that are not on the boundary of any dyadic cube. The *dyadic distance* of two points $x, y \in Q_0$ is the size of the smallest dyadic cube that contains both x and y. This defines a metric on Q_0. In a current work M. Csörnyei and P. Jones showed that:

Theorem 3.3. *(i) For any set $M \subset Q_0 \subset \mathbb{R}^n$ having m^n points there is a function $\psi\colon \mathbb{R}^{n-1} \to \mathbb{R}$ with dyadic Lipschitz constant 1 and there is a coordinate-direction x_k ($k = 1, 2, \dots, n$) such that the set*

$$\{(x_1, \dots, x_n) \in M : x_k = \psi(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)\}$$

has at least m^{n-1} points.

(ii) Every set $E \subset Q_0 \subset \mathbb{R}^n$ of Lebesgue measure m can be covered by dyadic Lipschitz slabs of total width at most $m^{1/n}$.

References

[1] G. Alberti, *Rank one property for derivatives of functions with bounded variation*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), 239–274.

[2] G. Alberti, M. Csörnyei and D. Preiss, paper in preparation.

[3] D.N. Bessis and F. H. Clarke, *Partial subdifferentials, derivates and Rademacher's theorem*, Trans. Amer. Math. Soc. 351 (1999), no. 7, 2899–2926.

[4] R. O. Davies, *On accessibility of plane sets and differentiation of functions of two real variables*, Proc. Cambridge Philos. Soc., 48 (1952), 215–232.

[5] M. Doré and O. Maleva, *A universal differentiability set in Banach spaces with separable dual*, in preparation

[6] M. Doré and O. Maleva, *Fréchet-differentiability of planar valued Lipschitz functions on Hilbert spaces*, in preparation

[7] T. Fowler and D. Preiss, *A simple proof of Zahorski's description of non-differentiability sets of Lipschitz functions,* Real. Anal. Exchange, 34 (2008/2009), no. 1, 1–12.

[8] J. Lindenstrauss, D. Preiss and J. Tišer, *Fréchet differentiability of Lipschitz functions and porous sets in Banach spaces*, in preparation

[9] D. Preiss, *Differentiability of Lipschitz functions on Banach spaces*, J. Funct. Anal. 91 (1990), no. 2, 312–345.

[10] N. X. Uy, *Removable sets of analytic functions satisfying a Lipschitz condition*, Ark. Mat. 17 (1979), no. 1, 19–27.

[11] Z. Zahorski, *Sur lensemble des points de non-derivabilite dune fonction continue*, Bull. Soc. Math. France 74 (1946), 147–178.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Asymptotic Analysis of the Toeplitz and Hankel Determinants via the Riemann-Hilbert Method

Alexander R. Its*

Abstract

The basic features of the asymptotic analysis of Toeplitz and Hankel determinants via the Riemann-Hilbert method including the fundamental connections to the theory of Painlevé equations are outlined. Some of the most recent results obtained in the field are discussed.

Mathematics Subject Classification (2010). Primary 47B35, 15B52; Secondary 35Q15, 34M55 .

Keywords. Toeplitz determinants, Riemann-Hilbert problem, Painlevé equations

1. Introduction

Let $\phi(z)$ be a function on the unit circle,

$$C = \{z : |z| = 1\}.$$

The *Toeplitz determinant,* $D_n^T[\phi]$, is defined as

$$D_n^T[\phi] := \det T_n[\phi], \tag{1}$$

where

$$T_n[\phi] := \{\phi_{j-k}\}, \quad k = 0, ..., n-1,$$

and

$$\phi_k = \int_C \phi(z) z^{-k-1} \frac{dz}{2\pi i}. \tag{2}$$

*Indiana University Purdue University Indianapolis, Department of Mathematical Sciences, 402 North Blackford Street, Indianapolis, Indiana, 46202-3216, USA.
E-mail: itsa@math.iupui.edu.

Similarly, given a function $w(x)$ on the real line $\mathbf{R}$ the *Hankel determinant*, $D_n^H[w]$, is defined as

$$D_n^H[w] := \det H_n[w], \tag{3}$$

where

$$H_n[\phi] := \{w_{j+k}\}, \quad k = 0, ..., n-1,$$

and

$$w_k = \int_{-\infty}^{\infty} x^k w(x) dx. \tag{4}$$

The principal analytic question is evaluation of the large n asymptotics of D_n^T and D_n^H.

Starting with Onsager's celebrated solution of the two-dimensional Ising model in the 1940's, Toeplitz and Hankel determinants play an increasingly central role in modern mathematical physics. Simultaneously, the theory of Toeplitz and Hankel determinants is a very beautiful area of analysis representing an unusual combinations of profound general operator concepts with the highly nontrivial concrete formulae. The area has been thriving since the classical works of Szegő, Fisher, Hartwig, Lenard and Widom, and it very much continious to do so.

In the 90s, it has been realized ([21], [1], [13]) that the theory of Toeplitz and Hankel determinants can be also embedded in the Riemann-Hilbert formalism of integrable systems. The new Riemann-Hilbert techniques have gone far beyond the classical Wiener-Hopf schemes, and they have led to the solutions of several important long-standing asymptotic problems of the theory. We shall review some of the most recent results which includes the proof of the Basor-Tracy conjecture concerning the asymptotics of Toeplitz determinants with the most general Fisher-Hartwig type symbols, the Fisher-Hartwig type asymptotics for Hankel determinants and for Toeplitz + Hankel determinants, and the transition asymptotics involving the Painlevé functions. The Riemann-Hilbert approach will be outlined as well.

The presentation is based on the author's joint works with P. Deift, T. Claeys, and I. Krasovsky.

2. Szegö and Fisher-Hartwig Asymptotics

The large n asymptotic behavior of $D_n^T[\phi]$ depends significantly on the analytical properties of the generating function $\phi(z)$. In the case of the smooth enough functions $\phi(z)$, the behavior is exponential and its leading and the pre-exponential terms are given by the following classical result of Szegő, known as the *strong Szegő limit theorem*.

Theorem 2.1. *Suppose that the generation function $\phi(z)$ satisfies the conditions,*

1. $\phi(z) \neq 0$ *, for all* $|z| = 1$.
2. index $\phi(z) \equiv \arg \phi(e^{2\pi i}) - \arg \phi(e^{i0}) = 0$
3. $\sum_{k=-\infty}^{\infty} |k||V_k|^2 < \infty$, *where* V_k *are the Fourier coefficients of the function,*

$$V(z) := \ln \phi(z), \tag{5}$$

that is,

$$V(z) = \sum_{k=-\infty}^{\infty} V_k z^k, \qquad V_k = \frac{1}{2\pi} \int_0^{2\pi} V(e^{i\theta}) e^{-ki\theta} d\theta. \tag{6}$$

Then,

$$D_n^T[\phi] \sim E_{\mathrm{Sz}}[\phi] \exp\big(nV_0\big), \quad n \to \infty, \tag{7}$$

where the pre-exponential factor, $E_{\mathrm{Sz}}[\phi]$, is given by the equation

$$E_{\mathrm{Sz}}[\phi] = \exp\left(\sum_{k=1}^{\infty} kV_kV_{-k}\right). \tag{8}$$

Conditions (1) and (2) on the symbol $\phi(z)$ ensure that the function $V(z)$ is a well defined function on the unit circle. Condition (3) is a smoothness condition which is, in fact, precise. In [33], Szegő proved this theorem under the assumption that the symbol is positive and that the symbol and its derivative are Lipshitz functions. It took a substantial period of time and the efforts of several very skillful analysts to reduce the smoothness conditions to the conditions (1) - (3) above. We refer the reader to the recent monograph of B. Simon [32] for all the history details.

Conditions of the strong Szegő theorem are not satisfied for the symbols having root and/or jump singularities. This type of symbols, in context of the statistical mechanics, was first considered in the works of M. Fisher and R. Hartwig [20], and A. Lenard [30]. These singularities are usually called the *Fisher-Hartwig singularities*. The general form of the symbol $\phi(z)$ which has m, $m = 0, 1, 2, \ldots$ fixed Fisher-Hartwig singularities can be given by the equation

$$\phi(z) = e^{V(z)} z^{\sum_{j=0}^{m} \beta_j} \prod_{j=0}^{m} |z - z_j|^{2\alpha_j} g_{z_j,\beta_j}(z) z_j^{-\beta_j}, \qquad z = e^{i\theta}, \qquad \theta \in [0, 2\pi), \tag{9}$$

where

$$z_j = e^{i\theta_j}, \quad j = 0, \ldots, m, \qquad 0 = \theta_0 < \theta_1 < \cdots < \theta_m < 2\pi; \tag{10}$$

$$g_{z_j,\beta_j}(z) \equiv g_{\beta_j}(z) = \begin{cases} e^{i\pi\beta_j} & 0 \leq \arg z < \theta_j, \\ e^{-i\pi\beta_j} & \theta_j \leq \arg z < 2\pi \end{cases}, \tag{11}$$

$$\Re\alpha_j > -1/2, \quad \beta_j \in \mathbb{C}, \quad j = 0, \ldots, m, \tag{12}$$

and $V(z)$ is a sufficiently smooth function on the unit circle so that the first factor of the right hand side of equation (9) represents the " Szegő part" of the symbol. The presence of the roots and jumps yield the appearance of the power-like factors in the large n behavior of the Toeplitz determinant. Indeed, the formula for the asymptotics of D_n^T now reads

$$D_n^T[\phi] \sim E_{\rm FH}[\phi] n^{\sum_{j=0}^m (\alpha_j^2-\beta_j^2)} \exp\Big(nV_0\Big), \quad n\to\infty. \tag{13}$$

The pre-exponential constant factor, $E_{\rm FH}[\phi]$, is more elaborated than its Szegő counterpart $E_{\rm Sz}[\phi]$ from the Szegő equation (7). The description of $E_{\rm FH}[\phi]$ involves a rather "exotic" special function - the Barnes' G - function $G(x)$ which is defined by the equations (see e.g. [36]),

$$G(1+x) = (2\pi)^{x/2} e^{-(x+1)x/2-\gamma x^2/2} \prod_{n=1}^{\infty} \{(1+x/n)^n e^{-x+x^2/(2n)}\}, \tag{14}$$

where γ is Euler constant. In addition to the Barnes' G - function, the formula for $E_{\rm FH}[\phi]$ involves the canonical Wiener-Hopf factorization of the Szegő part, $e^{V(z)}$, of the symbol $\phi(z)$,

$$e^{V(z)} = b_+(z) e^{V_0} b_-(z), \qquad b_+(z) = e^{\sum_{k=1}^{\infty} V_k z^k}, \qquad b_-(z) = e^{\sum_{k=-\infty}^{-1} V_k z^k}. \tag{15}$$

Note that $b_+(z)$ and $b_-(z)$ are analytic inside and outside of the unit circle $|z|=1$, respectively, and they satisfy the normalization conditions $b_+(0) = b_-(\infty) = 1$. The exact expression for $E_{\rm FH}[\phi]$ is given by the equation (cf. (8)),

$$\begin{aligned} E_{\rm FH}[\phi] = \exp\left(\sum_{k=1}^{\infty} kV_kV_{-k}\right) \prod_{j=0}^{m} b_+(z_j)^{-\alpha_j+\beta_j} b_-(z_j)^{-\alpha_j-\beta_j} \\ \times \prod_{0\le j<k\le m} |z_j-z_k|^{2(\beta_j\beta_k-\alpha_j\alpha_k)} \left(\frac{z_k}{z_j e^{i\pi}}\right)^{\alpha_j\beta_k-\alpha_k\beta_j} \\ \times \prod_{j=0}^{m} \frac{G(1+\alpha_j+\beta_j)G(1+\alpha_j-\beta_j)}{G(1+2\alpha_j)}. \end{aligned} \tag{16}$$

Asymptotics (13) was conjectured by M. Fisher and R. Hartwig in 1968 [20]. In the case of the root singularities only (all β are zero) formulae (13) - (16) were proven by H. Widom in 1973 [35]. The proof of the formulae (13) - (16) in the presence of jumps is due to Basor [4] for $\Re\beta_j = 0$, Böttcher and Silbermann [12] for $|\Re\alpha_j| < 1/2$, $|\Re\beta_j| < 1/2$, and Ehrhardt [18] for the only restriction, $|\Re\beta_j - \Re\beta_k| < 1$. We refer to [18] for a detail review of these and other related results. The precise statement concerning the large n behavior of the Toeplitz determinant $D_n^T[\phi]$ with the Fisher-Hartweg generating function (9) is given by the following theorem, proven by Ehrhardt [18].

Theorem 2.2. *([18]) Let $\phi(z)$ be defined in (9), $V(z)$ be C^∞ on the unit circle, $\Re\alpha_j > -1/2$, $|\Re\beta_j - \Re\beta_k| < 1$, and $\alpha_j \pm \beta_j \neq -1, -2, \ldots$ for $j, k = 0, 1, \ldots, m$. Then, as $n \to \infty$, the asymptotic behavior of the Toeplitz determinant $D_n^T[\phi]$ is given by the formulae (13) - (16).*

The condition,

$$|\Re\beta_j - \Re\beta_k| < 1, \quad \forall j, k = 0, 1, ..., m, \tag{17}$$

is precise. Indeed, A. Böttcher and B. Silbermann [12] in 1985 and E. Basor and C. Tracy [8] in 1991 found the examples with $\Re\beta_j$ not lying in a single interval of length less than 1 and such that the large n asymptotics is very different from the one given by (13). In the case of arbitrary complex β_j, E. Basor and C. Tracy conjectured in [8] the following description of the large n behavior of the determinant $D_n^T[\phi]$.

Let $\phi(z; n_0, ..., n_m)$ be a *representation* of the Fisher-Hartwig symbol $\phi(z)$ (9) defined by the equation,

$$\phi(z; n_0, ..., n_m) := \phi(z)|_{\beta_j \to \beta_j + n_j}, \quad \sum n_j = 0. \tag{18}$$

We note that, all representations of $\phi(z)$ differ only by multiplicative constants,

$$\phi(z) = \prod_{j=0}^{m} z_j^{n_j} \times \phi(z; n_0, \ldots, n_m). \tag{19}$$

Among all the representations $\phi(z; n_0, ..., n_m)$ of the symbol $\phi(z)$, we single out the set,

$$\mathcal{M} = \{\phi(z; n_0, ..., n_m) : \sum_{j=0}^{m} (\Re\beta_j + n_j)^2 \text{ is minimal}\}. \tag{20}$$

Theorem 2.3. *(Basor - Tracy conjecture) Let $\phi(z)$ be given in (9), $\Re\alpha_j > -1/2$, $\beta_j \in \mathbb{C}$, $j = 0, 1, \ldots, m$. Let $\mathcal{M}$ be non-degenerate, i.e. it contains no representations such that $\alpha_j + (\beta_j + n_j)$ or $\alpha_j - (\beta_j + n_j)$ is a negative integer for some j. Then, as $n \to \infty$,*

$$D_n^T[\phi] = \sum \left(\prod_{j=0}^{m} z_j^{n_j} \right)^n \mathcal{R}(\phi(z; n_0, \ldots, n_m))(1 + o(1)), \tag{21}$$

where the sum is over all representations in $\mathcal{M}$. Each $\mathcal{R}(\phi(z; n_0, \ldots, n_m))$ stands for the right-hand side of the formula (13), corresponding to $\phi(z; n_0, \ldots, n_m)$.

In the case of unique minimizer, this theorem was proven by T. Ehrhardt in 2001 [18]. The general case of the Basor-Tracy conjecture has been proven in [15] with the help of the new technique - the *Riemann-Hilbert method*.

3. The Riemann-Hilbert Method

The Riemann-Hilbert approach to the Toeplitz determinants was first suggested in [1] (see also [13]) as an extension to the Toeplitz case of the similar approach introduced earlier in [21] for Hankel determinants. The method is based on the classical relation between the Toeplitz determinants and the orthogonal polynomials on the unite circle.

Let us define the polynomials, $p_k(z) = \chi_k z^k + \cdots$, $\widehat{p}_k(z) = \chi_k z^k + \cdots$ of degree k, satisfying

$$\frac{1}{2\pi i}\int_0^{2\pi i} p_k(z) z^{-j}\phi(z)\frac{dz}{z} = \chi_k^{-1}\delta_{jk}, \qquad \frac{1}{2\pi i}\int_0^{2\pi} \widehat{p}_k(z^{-1}) z^{j}\phi(z)\frac{dz}{z} = \chi_k^{-1}\delta_{jk}, \tag{22}$$
$$j = 0, 1, \ldots, k.$$

Note, that relations (22) are equivalent to the equation,

$$\frac{1}{2\pi i}\int_0^{2\pi i} p_k(z)\widehat{p}_k(z^{-1})\phi(z)\frac{dz}{z} = \delta_{jk}. \tag{23}$$

In the case of real valued symbol $\phi(z)$, we obviously have that $\widehat{p}_k(z) = \overline{p_k(\bar{z})}$ and $p_k(z)$ becomes the orthogonal polynomials on the unite circle with respect to the *weight* $\phi(z)$. Assuming that the polynomials $p_k(z)$, $\widehat{p}_k(z)$ exist (which is, in particular, always the case for positive $\phi(z)$) the following general formula connecting $p_k(z)$ and the Toeplitz determinant $D_n^T[\phi]$ is valid (see e.g. [32]),

$$\frac{D_{n+1}^T}{D_n^T} = \chi_n^{-2}. \tag{24}$$

Formula (24) reduces the asymptotic analysis of the Toeplitz determinant $D_n[\phi]$ to the asymptotic analysis of the orthogonal polynomials $p_k(z)$ which in turn can be translated to the asymptotic analysis of the following matrix Riemann-Hilbert problem posed on the counterclockwise oriented unite circle $C = \{z : |z| = 1\}$ for a 2×2 matrix valued function $Y(z)$.

(a) $Y(z)$ is analytic for $z \in \mathbb{C} \setminus C$.

(b) Let $z \in C \setminus \cup_{j=0}^m z_j$. Y has continuous boundary values $Y_+(z)$ as z approaches the unit circle from the inside, and $Y_-(z)$, from the outside, related by the jump condition

$$Y_+(z) = Y_-(z)\begin{pmatrix} 1 & z^{-k}\phi(z) \\ 0 & 1 \end{pmatrix}, \qquad z \in C \setminus \cup_{j=0}^m z_j. \tag{25}$$

(c) $Y(z)$ has the following asymptotic behavior at infinity:

$$Y(z) = \left(I + O\left(\frac{1}{z}\right)\right)\begin{pmatrix} z^k & 0 \\ 0 & z^{-k} \end{pmatrix}, \qquad \text{as } z \to \infty. \tag{26}$$

(d) As $z \to z_j$, $j = 1, \dots, m$, $z \in \mathbb{C} \setminus C$,

$$Y(z) = \begin{pmatrix} O(1) & O(1) + O(|z - z_j|^{2\alpha_j}) \\ O(1) & O(1) + O(|z - z_j|^{2\alpha_j}) \end{pmatrix}, \qquad \text{if } \alpha_j \neq 0, \tag{27}$$

and

$$Y(z) = \begin{pmatrix} O(1) & O(\ln|z - z_j|) \\ O(1) & O(\ln|z - z_j|) \end{pmatrix}, \qquad \text{if } \alpha_j = 0,\ \beta_j \neq 0. \tag{28}$$

Having the solution $Y(z)$ of the Riemann-Hilbert problem, the orthogonal polynomials $p_k(z)$, $\widehat{p}_k(z)$ and the coefficient χ_k can be reconstructed with the help of the equations,

$$\chi_{k-1}^2 = -Y_{21}(0), \quad p_k(z) = \chi_k Y_{11}(z), \quad \widehat{p}_k(z) = -\frac{z^{k-1}}{\chi_{k-1}} Y_{21}(z^{-1}). \tag{29}$$

The asymptotic analysis of the above formulated Riemann-Hilbert problem can be performed with the help of the *nonlinear steepest descent method.* This method was introduced in the theory of integrable systems in 1992 by Deift and Zhou [17] and extended to the Riemann-Hilbert problems appearing in the orthogonal polynomial theory in the work [16]. The methodological idea of the method is to perform a sequence of exact deformations of the given jump - matrices and contours in order to transform the Riemann-Hilbert problem in question to an equivalent Riemann-Hilbert problem whose jump matrices are uniformly close to identity. As a result of these deformations, the asymptotic solution of the original Riemann-Hilbert problem reduces to the solution of certain model local Riemann-Hilbert problems associated with the special points and domains arising in the course of the deformation. The noncommutativity of the matrix setting requires, however, the development of several rather sophisticated new technical ideas which, in particular, enable an explicit solution of the model Riemann-Hilbert problems. The final result of the analysis is as efficient as the asymptotic evaluation of the oscillatory integrals. For more detail we refer to the original papers, [17], [16],[1], [15], to the lecture notes [23], and to the monograph [14].

The important advantage of using the Riemann-Hilbert formalism for the asymptotic evaluation of the Toeplitz determinants is that the method also provides the asymptotics of the corresponding orthogonal polynomials. The orthogonal polynomials are involved in many important algebraic identities of the theory of Toeplitz as well as Hankel determinants. The possibility of the asymptotic evaluation of $p_k(z)$ and $\widehat{p}_k(z)$ allow to use these identities in the situations where the direct analysis of D_n is not available. In particular, the proof of the Basor-Tracy conjecture obtained in [15] is based on the combination of the asymptotic analysis of the Riemann-Hilbert problem[1] (a) - (d) with the

[1]The asymptotic analysis of the Riemann-Hilbert problem (a) - (d), in the case when the only root singularities are present was also (and earlier) considered in [31]

following general identity.

$$D_n^T[z^l\phi(z)] = (-1)^{ln}\frac{F_n}{\prod_{j=1}^{l-1} j!}D_n^T[\phi(z)], \quad l \in \mathbb{Z}_+, \tag{30}$$

where

$$F_n = \det\{P_{n+j}^{(k)}(0)\}_{j,k=0,\dots l-1}, \tag{31}$$

and

$$P_k(z) = p_k(z)/\chi_k \equiv Y_{11}(z). \tag{32}$$

This identity expresses the Toeplitz determinant with the symbol $z^l\phi(z)$ in terms of the original Toeplitz determinant $D_n^T[\phi]$ and the orthogonal polynomials $p_k(z)$. The asymptotics of the latter are provided via the asymptotic analysis of the Riemann-Hilbert problem.

4. Fisher-Hartwig Asymptotics for Hankel Determinants

Consider the Hankel determinants $D_n^H[w]$ whose symbols are supported on the finite interval $[-1,1]$. The following equation establishes a direct link of this type of Hankel determinants with the Toeplitz determinants.

$$\left(D_n^H[w(x)]\right)^2 = \frac{\pi^{2n}}{4^{(n-1)^2}}\frac{(1+P_{2n}(0))^2}{P_{2n}(1)P_{2n}(-1)}D_{2n}^T[\phi(z)], \tag{33}$$

where the symbols $w(x)$ and $\phi(z)$ are related by the equation,

$$w(x) = \frac{\phi(e^{i\theta})}{|\sin\theta|}, \qquad x = \cos\theta, \quad x \in [-1,1], \tag{34}$$

and $P_k(z)$ denote the monic orthogonal polynomials (32) on the unit circle with the weight $\phi(z)$.

Suppose that the symbol $w(x)$ is of general Fisher-Hartwig form, i.e.

$$w(x) = e^{U(x)}\prod_{j=0}^{r+1}|x-\lambda_j|^{2\alpha_j}\omega_j(x), \quad 1 = \lambda_0 > \lambda_1 > \cdots > \lambda_{r+1} = -1,$$

$$\omega_j(x) = \begin{cases} e^{i\pi\beta_j} & \Re x \le \lambda_j \\ e^{-i\pi\beta_j} & \Re x > \lambda_j \end{cases}, \qquad \Re\beta_j \in (-1/2, 1/2],$$

$$\beta_0 = \beta_{r+1} = 0, \qquad \Re\alpha_j > -\frac{1}{2}, \qquad j = 0,1,\dots,r+1. \tag{35}$$

where $U(x)$ is a sufficiently smooth function on the interval $[-1,1]$. Identity (33) allows to obtain the asymptotics of the Hankel determinants with the

Fisher-Hartwig symbols (35) from the similar asymptotics of the Toeplitz determinants provided the asymptotics of the related orthogonal polynomials on the unit circle are known. The orthogonal polynomial asymptotics are available via the Riemann-Hilbert approach. The exact statement concerning the leading asymptotics of the Hankel determinant $D_n^H[w]$ reads as following.

Theorem 4.1 ([15]). *Let $w(x)$ be defined as in (35) with $\Re\beta_j \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, $j = 1, 2, \ldots, r$. Then as $n \to \infty$,*

$$D_n^H[w]/D_n^H[1] \sim E_{FH}^H[w] n^{2(\alpha_0^2+\alpha_{r+1}^2)+\sum_{j=1}^r(\alpha_j^2-\beta_j^2)}$$

$$\times \exp\left(nV_0 + 2in\sum_{j=1}^r \beta_j \arcsin\lambda_j - 2n\sum_{j=0}^{r+1}\alpha_j \ln 2\right), \tag{36}$$

where $V(e^{i\theta}) = U(\cos\theta)$, and the constant pre - factor $E_{FH}^H[w]$ admits an explicit representation[2] *in terms of the parameters of the weight $w(x)$ and the Wiener-Hopf functions $b_\pm(z)$ (15).*

The above results can be extended to the case of the symbols $w(x)$ supported on the infinite interval as well. Formula (33) is not applicable in this case, of course. However, one can proceed with the direct Riemann-Hilbert analysis of Hankel determinants using the Riemann-Hilbert representation of the orthogonal polynomials on the line [21]. We refer to the work [24] and to the references therein for the results concerning the asymptotics of Hankel determinants with the symbols on the line and having the Fisher-Hartwig singularities.

5. Toeplitz + Hankel Determinants. The *L* - functions

Let $\phi(z)$ be the Fisher-Hartwig symbol defined by equations (9) and, as before, denote ϕ_k its Fourier coefficients (2). The so-called Toeplitz + Hankel determinants are defined as the following three types of the determinants,

$$\det(\phi_{j-k} + \phi_{j+k})_{j,k=0}^{n-1}, \quad \det(\phi_{j-k} - \phi_{j+k+2})_{j,k=0}^{n-1}, \quad \det(\phi_{j-k} \pm \phi_{j+k+1})_{j,k=0}^{n-1}. \tag{37}$$

These determinants appear in the theory of classical groups and its applications to random matrices, statistical mechanics, and number theory (see, e.g., [2, 19, 29, 10]). In all applications mentioned, the symbol $\phi(z)$ is an even function on the circle, i.e. it satisfies the addition symmetry, $\phi(e^{-i\theta}) = \phi(e^{i\theta})$, which implies that the matrices in (37) are symmetric.

[2]This representation is similar though even more involved to the representation (16) for the Toeplitz pre-factor $E_{FH}[\phi]$. For the exact formula for $E_{FH}^H[w]$ see [15].

The key observation [34, 25, 2] (see also [15], Lemma 2.7) is that there are simple relations between the determinants (37) and Hankel determinants on $[-1,1]$ with added singularities at the end-points. Indeed, the following formulae take place,

$$\det(\phi_{j-k}+\phi_{j+k})_{j,k=0}^{n-1}=\frac{2^{n^2-2n+2}}{\pi^n}D_n\Big(\phi(e^{i\theta(x)})/\sqrt{1-x^2}\Big) \tag{38}$$

$$\det(\phi_{j-k}-\phi_{j+k+2})_{j,k=0}^{n-1}=\frac{2^{n^2}}{\pi^n}D_n\Big(\phi(e^{i\theta(x)})\sqrt{1-x^2}\Big), \tag{39}$$

$$\det(\phi_{j-k}+\phi_{j+k+1})_{j,k=0}^{n-1}=\frac{2^{n^2-n}}{\pi^n}D_n\Big(\phi(e^{i\theta(x)})\sqrt{\frac{1+x}{1-x}}\Big), \tag{40}$$

$$\det(\phi_{j-k}-\phi_{j+k+1})_{j,k=0}^{n-1}=\frac{2^{n^2-n}}{\pi^n}D_n\Big(\phi(e^{i\theta(x)})\sqrt{\frac{1-x}{1+x}}\Big), \tag{41}$$

where, as before, $x=\cos\theta$. These formulae in conjunction with Theorem 4.1 yield immediately the asymptotic expansions for Toeplitz + Hankel determinants (37) with even Fisher-Hartwig symbols (9) whose $\Re\beta_j\in\left(-\frac{1}{2},\frac{1}{2}\right)$. The exact statement contains in Theorem 1.25 of [15].

Remark 5.1. Earlier, using the direct operator-theoretical methods, the asymptotics of the determinant $\det(\phi_{j-k}+\phi_{j+k+1})_{j,k=0}^{n-1}$, in the case of all $\alpha=0$, was obtained in [5] and in the case of non-even ϕ and still all $\alpha=0$ in [6]. Recently, for smooth symbols, the asymptotics of all the determinants (37) (and related more general ones) were found in [7].

The asymptotic formulae for the Toeplitz+Hankel determinants obtained in [15] have an interesting application arising in the framework of the random matrix approach in the theory of the Riemann zeta - function and other L - functions [28]. Define

$$\phi(z)=\left|2\sin\frac{\theta}{2}\right|^{2k}e^{V(z)},\quad k\in\mathbb{N}, \tag{42}$$

where

$$V(e^{i\theta})=2k\left\{\int_1^e u(y)\left(\sum_{j=-\infty}^{\infty}\mathrm{Ci}(|\theta+2\pi j|\ln y\ln X)dy\right)-\ln\left|2\sin\frac{\theta}{2}\right|\right\},$$

$$\mathrm{Ci}(z)=-\int_z^\infty\frac{\cos t}{t}dt,$$

and $u(y)$ is a smooth nonnegative function supported on $[e^{1-1/X},e]$ and of total mass one. Consider the following average over the orthogonal group $SO(2n)$,

$$E_{SO(2n)}\left(\prod_{j=1}^{n}\phi(e^{i\theta_j})\right). \tag{43}$$

We are interested in the large n and large X behavior of this average. Observe that

$$E_{SO(2n)}\left(\prod_{j=1}^{n} f(e^{i\theta_j})\right) = \frac{1}{2}\det(\phi_{j-k}+\phi_{j+k})_{j,k=0}^{n-1},$$

and that symbol (42) is of Fisher-Hartwig type with a single α - singularity at $z_0 = 1$, and $\alpha_0 = k$. A direct application of Theorem 1.25 of [15] leads then to the following asymptotic behavior of average (43),

$$E_{SO(2n)}\left(\prod_{j=1}^{n} f(e^{i\theta_j})\right) \sim G(1+k)\left(\frac{\Gamma(1+2k)}{G(1+2k)\Gamma(1+k)}\right)^{1/2}\left(\frac{2n}{e^{\gamma}\ln X}\right)^{k(k-1)/2}, \tag{44}$$

where γ is Euler's constant. Formula (44) has been already conjectured by Bui and Keating as the random matrix counterpart of a relevant number theoretical conjecture concerning the mean values of certain Dirichlet L - functions in the Katz-Sarnak orthogonal family [10]. In a similar way, Theorem 1.25 of [15] provides a justification of the Bui-Keating conjecture about the average of the same product $\prod_{j=1}^{n}\phi(e^{i\theta_j})$ over the symplectic group. In context of the random matrix approach in number theory, this means that the asymptotic results of [15] support the number theoretical conjectures of [10].

6. Transition Asymptotics and Painlevé Functions

Painlevé transcendents appear in the framework of the Riemann-Hilbert method in a very natural way. Indeed, they represent the model Riemann-Hilbert problems associated with the paramterices to the solution of the original Riemann-Hilbert problem at its coalescing special points. The situation is very similar to the appearance of the linear counterparts of the Painlevé functions, i.e. Airy functions, Bessel functions, etc., in the asymptotic analysis of an oscillatory contour integral when its stationary points or/and poles coalesce.

Consider the one-parameter family of Toeplitz determinants $D_n(t)$ with the symbol,

$$\phi(z) \equiv \phi(z;t) = (z-e^{t})^{\alpha+\beta}(z-e^{-t})^{\alpha-\beta}z^{-\alpha+\beta}e^{-i\pi(\alpha+\beta)}e^{V(z)}, \tag{45}$$

$$\alpha \pm \beta \neq -1, -2, ...,$$

where $t \geq 0$ is sufficiently small and $\alpha, \beta \in \mathbb{C}$ with $\Re\alpha > -\frac{1}{2}$. The potential $V(z)$ is assumed to be analytic in an annulus containing the unit circle. When $t > 0$, the symbol (45) is smooth (in fact, analytic) on the unit circle and the Toeplitz determinant exhibits the Szegő type assymptotic behavior. When $t = 0$, the

branch points coalesce at $z = 1$ and the symbol becomes of the Fisher-Hartwig type. The asymptotics of $D_n(t)$ transforms to the Fisher-Hartwig asymptotics. The analysis of the corresponding Rimeann-Hilbert problem [11] shows that the transition asymptotics, which is uniform with respect to $t \geq 0$, is given in terms of a special solution to the fifth Painlevé equation. We shall now present the exact formulation of this result for the case (for general case - see [11]).

$$\Re\beta = 0, \quad \alpha > -\frac{1}{2} \in \mathbb{R}, \quad 2\alpha \notin \mathbb{Z}. \tag{46}$$

Consider the Jimbo-Miwa-Okamoto σ - form of the fifth Painlevé equation (cf. [26]),

$$\left(x\frac{d^2\sigma}{dx^2}\right)^2 = \left(\sigma - x\frac{d\sigma}{dx} + 2\left(\frac{d\sigma}{dx}\right)^2 + 2\alpha\frac{d\sigma}{dx}\right)^2$$

$$-4\left(\frac{d\sigma}{dx}\right)^2\left(\frac{d\sigma}{dx} + \alpha + \beta\right)\left(\frac{d\sigma}{dx} + \alpha - \beta\right). \tag{47}$$

The particular choice of the solution to this equation which is participating in the uniform asymptotic expansion of the determinant $D_n(t)$ is characterized by the following asymptotic condition as $x \to 0$,

$$\sigma(x) = \alpha^2 - \beta^2 + \frac{\alpha^2 - \beta^2}{2\alpha}x\Big(1 - x^{2\alpha}C(\alpha,\beta)\Big)(1 + O(x)), \tag{48}$$

where

$$C(\alpha,\beta) = \frac{\Gamma(1+\alpha+\beta)\Gamma(1+\alpha-\beta)}{\Gamma(1-\alpha+\beta)\Gamma(1-\alpha-\beta)}\frac{\Gamma(1-2\alpha)}{\Gamma(1+2\alpha)^2}\frac{1}{1+2\alpha}.$$

Under the assumptions (46), the solution $\sigma(x)$ is real and has no singularities for $x > 0$. As $x \to +\infty$, the function $\sigma(x)$ decays exponentially. The leading asymptotic behavior of the solution $\sigma(x)$ for large positive x is given by the formulae,

$$\sigma(x) = -x^{-1+2\alpha}e^{-x}\frac{1}{\Gamma(\alpha-\beta)\Gamma(\alpha+\beta)}\left(1 + O\left(\frac{1}{x}\right)\right), \quad x \to \infty. \tag{49}$$

The asymptotic expansion of the Toeplitz determinant with symbol (45) which interpolates between Szegő and Fisher-Hartwig asymptotics is given explicitly in terms of the function $\sigma(x)$ and reads as follows.

Theorem 6.1 ([11]). *Let α and β satisfy conditions (46), and let $\sigma(x)$ be the unique solution of the fifth Painlevé equation (47) characterized by either asymptotics (48) at $x = 0$, or by asymptotics (49) at $x = +\infty$. Then the*

following asymptotic expansion holds for the Toeplitz determinant $D_n(t)$ as $n \to \infty$ with the error term $O(1/n)$ uniform for $0 \le t < t_0$ where t_0 is sufficiently small:

$$\ln D_n(t) = nV_0 + (\alpha+\beta)nt + \sum_{k=1}^{\infty} k \left[V_k - (\alpha+\beta)\frac{e^{-tk}}{k}\right]\left[V_{-k} - (\alpha-\beta)\frac{e^{-tk}}{k}\right] + \ln \frac{G(1+\alpha+\beta)G(1+\alpha-\beta)}{G(1+2\alpha)} + \Omega(2nt) + O(1/n), \quad (50)$$

where $G(z)$ is Barnes' G-function (14) and

$$\Omega(2nt) = \int_0^{2nt} \frac{\sigma(x) - \alpha^2 + \beta^2}{x} dx + (\alpha^2 - \beta^2) \ln 2nt. \quad (51)$$

Taking into account the asymptotic properties of the Painlevé function $\sigma(x)$ and the identity,

$$\sum_{k=1}^{\infty} \frac{e^{-2tk}}{k} = -\ln(1 - e^{-2t}),$$

we see that the asymptotics (50) is the Szegö type asymptotics (7) for any $t > 0$, and it becomes the Fisher-Hartwig type asymptotics (13) when $t = 0$.

The statement of Theorem 6.1 remains essentially valid in general case of $\alpha, \beta \in \mathbb{C}$, $\Re\alpha > -\frac{1}{2}$. One only has to account for the possible finite number of real poles of the function $\sigma(x)$ (and real zeros of $D_n(t)$) and modify the formulation accordingly (see [11]). With these modifications, Theorem 6.1 generalizes the classical result of Wu-McCoy-Tracy-Barouch [37] concerning the Painlevé III - description of the phase transition in the large distance behavior of the 2-spin correlation function in the 2D Ising model. Indeed, the Ising phase transition corresponds to the Toeplitz determinant with the symbol of type (45) and the choice of the parameters $\alpha = 0$ ad $\beta = -\frac{1}{2}$. The Painlevé V equation (47), as it was shown in [27], can be then reduced to the third Painlevé equation.

The reason for appearance of the Painlevé function in expansion (50) is the coalescence of the branch points of symbol (45) on the unite circle as parameter t approaches zero. This yields the necessity to introduce in the neighborhood of these points a model Riemann-Hilbert problem whose relevant linear system [3] has two regular singular points and one irregular singular point of Poincaré rank 1. This Riemann-Hilbert problem is known to be the Riemann-Hilbert problem for the fifth Painlevé equation (see [26]; see also [22]).

[3] See e.g.[22] for a detail exposition of the classical connection of the Riemann-Hilbert problems and the monodromy theory of linear systems of differential equations with rational coefficients

Similar to Theorem 6.1 results featuring other types of transition regimes in the asymptotic behavior of Toeplitz and Hankel determinants had been earlier obtained in the works [1], [3] and [9].

References

[1] J. Baik, P. Deift, K. Johansson, *On the distribution of the length of the longest increasing subsequence of random permutations*, J. Amer. Math. Soc. **12** (1999), 1119–1178.

[2] J. Baik and E. M. Rains, *Algebraic aspects of increasing subsequences,* Duke Math. J. **109** (2001), 1–65.

[3] J. Baik and E. M. Rains, *The asymptotics of monotone subsequences of involutions*, Duke Math. J. **109** (2001), 205–281.

[4] E. Basor, *Asymptotic formulas for Toeplitz determinants*, Trans. Amer. Math. Soc. **239** (1978), 33–65.

[5] E. L. Basor and T. Ehrhardt, *Asymptotic formulas for the determinants of symmetric Toeplitz plus Hankel matrices. Toeplitz matrices and singular integral equations (Pobershau, 2001),* 61–90, Oper. Theory Adv. Appl., 135, Birkhauser, Basel, 2002.

[6] E. L. Basor and T. Ehrhardt, *Asymptotic formulas for determinants of a sum of finite Toeplitz and Hankel matrices,* Math. Nachr. **228** (2001), 5–45.

[7] E. L. Basor and T. Ehrhardt, *Determinant computations for some classes of Toeplitz-Hankel matrices* [arXiv:0804.3073]

[8] E. L. Basor and C. A. Tracy, *The Fisher-Hartwig conjecture and generalizations*, Phys. A **177** (1991), 167–173.

[9] P. Bleher, A. Its, *Asymptotics of the partition function of random matrix model*, Annales de l'Institute Fourier, **55**, 6 (2005), 1943–2000

[10] H. M. Bui and J. P. Keating, *On the mean values of L-functions in orthogonal and symplectic families*, Proc. London Math. Soc. (3) **96** (2008) 335–366; J. P. Keating. Private communication.

[11] T. Claeys, A. Its, I. Krasovsky, *Emergence of a singularity for Toeplitz determinant and Painlevé V*, prprint, 2010.

[12] A. Böttcher and B. Silbermann, *Toeplitz matrices and determinants with Fisher-Hartwig symbols*, J. Funct. Anal. **63** (1985), 178–214.

[13] P. Deift, *Integrable operators*; in the book: *Differential operators and spectral theory*, 69–84, Amer. Math. Soc. Transl. Ser. 2, **189**, Amer. Math. Soc., Providence, RI, 1999.

[14] P. A. Deift, *Orthogonal Polynomials and Random Matrices: A Riemann-Hilbert Approach*, Courant Lecture Notes in Mathematics, **3**, CIMS, New York, 1999.

[15] P. Deift, A. Its, and I. Krasovsky, *Asymptotics of Toeplitz, Hankel, and Toeplitz+Hankel determinants with Fisher-Hartwig singularities*, preprint, arXiv:0905.0443v1 [math.FA]

[16] P. Deift, T. Kriecherbauer, K.T-R McLaughlin, S. Venakides, and X. Zhou, *Strong asymptotics of orthogonal polynomials with respect to exponential weights*, Comm. Pure Appl. Math., **52** (1999), 1491–1552.

[17] P. Deift and X. Zhou, *A steepest descent method for oscillatory Riemann-Hilbert problems. Asymptotics for the MKdV equation*, Ann. Math. **137** no. 2 (1993), 295–368.

[18] T. Ehrhardt, *A status report on the asymptotic behavior of Toeplitz determinants with Fisher-Hartwig singularities*, Operator Theory: Adv. Appl. **124** (2001), 217–241.

[19] P. J. Forrester, N. E. Frankel, *Applications and generalizations of Fisher-Hartwig asymptotics,* J. Math. Phys. **45** (2004), 2003–2028.

[20] M. E. Fisher, R. E. Hartwig. *Toeplitz determinants: Some applications, theorems, and conjectures.,* Advan. Chem. Phys. **15** (1968), 333–353.

[21] A. S. Fokas, A .R. Its and A. V. Kitaev, *The Isomonodromy Approach to Matrix Models in 2D Quantum Gravity*, Commun. Math. Phys., **147** (1992), 395–430 .

[22] A. Fokas, A. Its, A. Kapaev, V. Novokshenov, *Painlevé Transcendents: The Riemann-Hilbert Approach,* AMS Mathematical Surveys and Monographs, **128**, 2006

[23] A. R. Its, *Large N asymptotics in random matrices. The Riemann-Hilbert approach*, Lecture Notes, 2005 CRM Short Programme on: "Random Processes and Integrable Systems".

[24] A. Its and I. Krasovsky, *Hankel determinant and orthogonal polynomials for the Gaussian weight with a jump*, Contemp. Math., **458** (2008), 215–247.

[25] K. Johansson, *On random matrices from the compact classical groups,* Ann. of Math. (2) **145** (1997), no. 3, 519–545.

[26] M. Jimbo, T. Miwa, *Monodromy preserving deformation of linear ordinary differential equations with rational coefficients. II,* Physica D, **2** (1981), 407–448.

[27] M. Jimbo and T. Miwa, *Studies on holonomic quantum fields XVII,* Proc. Japan Acad., **56** A (1980), 405–410.

[28] J. P. Keating, *Random matrices and number theory*, in book: *Applications of random matrices in physics*, eds. E. Brźin, V. Kazakov, D. Serban, P. Wiegmann, A. Zabrodin, NATO Science Series II. Mathematics, Physics and Chemistry - Vol. 221, Springer, 2006

[29] J. P. Keating, F. Mezzadri, *Random matrix theory and entanglement in quantum spin chains,* Comm. Math. Phys. **252** (2004), 543–579.

[30] A. Lenard. *Momentum distribution in the ground state of the one-dimensional system of impenetrable bosons*, J. Math. Phys. **5** (1964) 930–943; A. Lenard *Some remarks on large Toeplitz determinants.* Pacific J. Math. **42** (1972), 137–145.

[31] A. Martínez-Finkelshtein, K. T.-R. McLaughlin, E. B. Saff, *Asymptotics of orthogonal polynomials with respect to an analytic weight with algebraic singularities on the circle,* Int. Math. Res. Not. 2006, Art. ID 91426, 43 pp.

[32] B. Simon, *Orthogonal polynomials on the unit circle*, AMS Colloquium Publications, 2005.

[33] G. Szegő, *Orthogonal polynomials,* AMS Colloquium Publ. **23**. New York: AMS 1959.

[34] H. Weyl, *The classical groups,* Princeton University Press, Princeton, 1946.

[35] H. Widom. *Toeplitz determinants with singular generating functions*, Amer. J. Math. **95** (1973), 333–383

[36] E. Whittaker, G. Watson, *A course of modern analysis,* Cambridge, 1969.

[37] T. T. Wu, B. M. McCoy, C. A. Tracy and E. Barouch, *Spin-spin correlation functions for the two-dimensional Ising model: Exact theory in the scaling region,* Phys. Rev., **B13** (1976), 316–374.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Regularity of the Inverse of a Sobolev Homeomorphism

Pekka Koskela*

Abstract

We give necessary and sufficient conditions for the inverse of a Sobolev homeomorphism to be a Sobolev homeomorphism and conditions under which the inverse is of bounded variation.

Mathematics Subject Classification (2010). Primary 30C65; Secondary 46E35.

Keywords. Sobolev mapping, bounded variation, homeomorphism, inverse, finite distortion

1. Planar Sobolev Mappings

Consider the usual Cantor ternary function u on the interval $(0,1)$. Then u is continuous, non-decreasing, constant on each complementary interval of the ternary Cantor set, and fails to be absolutely continuous. Now, let $g(x) = x + u(x)$ on $(0,1)$. Then also g fails to be absolutely continuous and hence g does not belong to the Sobolev class $W^{1,1}_{\mathrm{loc}}((0,1),\mathbb{R})$. On the other hand, the Lipschitz function $h = g^{-1}$ maps $(0,2)$ homeomorphically onto $(0,1)$. Thus, even the inverse of a Lipschitz homeomorphism h can fail to belong to $W^{1,1}_{\mathrm{loc}}$. If one analyzes the situation more carefully, one notices that the crucial thing here is that h' vanishes in a set of positive measure; if this were not the case, $h^{-1} = g$ would necessarily belong to $W^{1,1}_{\mathrm{loc}}((0,1),\mathbb{R})$.

In dimension two, the mapping $f(x_1, x_2) = (h(x_1), x_2)$, where h is as above, provides us with a Lipschitz homeomorphism whose inverse fails to belong to $W^{1,1}_{\mathrm{loc}}(\mathbb{R}^2,\mathbb{R}^2)$. Here $W^{1,1}_{\mathrm{loc}}(\Omega;\mathbb{R}^2)$ consists of all locally integrable mappings of Ω into $\mathbb{R}^2$ whose both component functions have locally integrable distributional derivatives. Notice that the Jacobian determinant $J_f(x)$ of the mapping f above

*The author was supported by the Academy of Finland grants 120972, 131477.

Department of Mathematics and Statistics, University of Jyväskylä, P.O.Box 35 (MaD), FI-40014 University of Jyväskylä, Finland. E-mail: pkoskela@maths.jyu.fi.

vanishes in a set of positive area. Our first result from [9] shows that this is the only situation where the inverse fails to belong to $W^{1,1}_{\mathrm{loc}}$.

Theorem 1.1. *Let $\Omega \subset \mathbb{R}^2$ be a domain. Suppose that $f \in W^{1,1}_{\mathrm{loc}}(\Omega, \mathbb{R}^2)$ is a homeomorphism and that $J_f(x) > 0$ for a.e. $x \in \Omega$. Then $f^{-1} \in W^{1,1}_{\mathrm{loc}}(f(\Omega), \mathbb{R}^2)$ and*

$$\int_{f(\Omega)} |Df^{-1}(y)|\, dy = \int_{\Omega} |Df(x)|\, dx.$$

Notice that the the above result does not assert that $J_{f^{-1}}(y) > 0$ almost everywhere. It is not difficult to give examples where this property fails under the assumptions above. The following stronger result from [9] gives a symmetric statement.

Theorem 1.2. *Let $\Omega \subset \mathbb{R}^2$ be a domain. Suppose that $f \in W^{1,1}_{\mathrm{loc}}(\Omega, \mathbb{R}^2)$ is a homeomorphism, and assume further that $Df(x)$ vanishes almost everywhere in the zero set of J_f. Then $f^{-1} \in W^{1,1}_{\mathrm{loc}}(f(\Omega), \mathbb{R}^2)$ and $Df^{-1}(y)$ vanishes almost everywhere in the zero set of $J_{f^{-1}}$. Moreover,*

$$\int_{f(\Omega)} |Df^{-1}(y)|\, dy = \int_{\Omega} |Df(x)|\, dx.$$

Recall that a homeomorphism $f \in W^{1,1}_{\mathrm{loc}}(\Omega, f(\Omega))$ is classically differentiable almost everywhere [17], see also [5]. Thus either $J_f(x) \geq 0$ almost everywhere or $J_f(x) \leq 0$ almost everywhere. For simplicity, let us assume from now on that $J_f(x) \geq 0$ almost everywhere. Under the assumptions of Theorem 1.2, the inequality

$$|Df(x)|^2 \leq K_f(x) J_f(x)$$

then holds a.e., where $1 \leq K(x) < \infty$ a.e. In fact, the optimal such function is obtained by setting $K_f(x) = |Df(x)|^2 / J_f(x)$ when $J_f(x) > 0$ and $K_f(x) = 1$ otherwise (assuming that Df vanishes a.e. in the zero set of J_f). For simplicity, we then say that f is a mapping (or homeomorphism) of finite distortion K_f (cf. [13]). The proof of Theorem 1.2 actually allows for a stronger formulation.

Theorem 1.3. *Let $\Omega \subset \mathbb{R}^2$ be a domain and let $f \in W^{1,1}_{\mathrm{loc}}(\Omega, \mathbb{R}^2)$ be a homeomorphism. Then $f^{-1} \in W^{1,1}_{\mathrm{loc}}(f(\Omega), \mathbb{R}^2)$ if and only if f has finite distortion, and if either of these conditions hold, then f^{-1} also has finite distortion.*

It is now natural to inquire if a suitable integrability condition on K_f would guarantee better regularity for the inverse of f. Our next result from [9],[2] gives an affirmative answer.

Theorem 1.4. *Let $\Omega \subset \mathbb{R}^2$ be a domain. Suppose that $f \in W^{1,1}_{\mathrm{loc}}(\Omega, \mathbb{R}^2)$ is a homeomorphism of finite distortion with $K_f \in L^1(\Omega)$. Then $f^{-1} \in W^{1,2}_{\mathrm{loc}}(f(\Omega), \mathbb{R}^2)$ and f^{-1} is a mapping of finite distortion. Moreover,*

$$\int_{f(\Omega)} |Df^{-1}(y)|^2\, dy = \int_{\Omega} K_f(x)\, dx.$$

The identity from Theorem 1.4 indicates that L^1-minimization problems for K_f are related to harmonic mappings. This is indeed the case, but it turns out to be more convenient to use another distortion function, defined using the Hilbert-Schmidt norm instead of the operator norm of $Df(x)$ [2], [11]. For convex Ω and suitable boundary values, there is a unique homeomorphic minimizer for the L^1-minimization for this distortion function. Moreover, this minimizer is smooth and its inverse is a harmonic mapping.

Based on the conclusions of Theorem 1.2 and Theorem 1.4, it would be natural to expect for an interpolation-type result, where the integrability of a power $0 < a < 1$ of K would result in the q-integrability for $|Df^{-1}|$ for some $1 < q(a) < 2$. This turns out not to be the case [9].

Theorem 1.5. *Let $0 < \delta < 1$. There is a homeomorphism $f : B(0,1) \to B(0,1)$ of finite distortion such that $f \in W^{1,1}(B(0,1), \mathbb{R}^2)$ and $K_f^{1-\delta} \in L^1(B(0,1))$, but $f^{-1} \notin W^{1,1+\delta}_{\mathrm{loc}}(B(0,1), \mathbb{R}^2)$.*

As Theorem 1.5 easily implies, the integrability of $K_f^{1-\delta}$ does not necessarily result in any better than $W^{1,1}$-regularity of f^{-1}, even when δ is small. One could still hope for some improvement under some a priori assumption on f. This turns out to be the case: given a homeomorphism $f \in W^{1,p}_{\mathrm{loc}}(\Omega, \mathbb{R}^2)$, $p > 1$, of finite distortion with $K_f^a \in L^1(\Omega)$, $0 < a \leq 1$, one always has $f^{-1} \in W^{1,q}_{\mathrm{loc}}(f(\Omega), \mathbb{R}^2)$, where $1 < q(p,a) \leq 2$, see [9]. In the special case when f is Lipschitz, one can take $q = a + 1$. There would be no additional gain at an L^q-scale from an exponent $a > 1$: simply notice that the Lipschitz homeomorphism

$$f(x) = x||x||^s$$

has a bounded K for each (large) $s > 0$ and $f^{-1} \notin W^{1,q}_{\mathrm{loc}}$ for $q = 2 + 2/s$.

It had been known for a long time that $K_f \in L^1(\Omega)$ and $f \in W^{1,2}_{\mathrm{loc}}(\Omega, \mathbb{R}^2)$ guarantee that $f^{-1} \in W^{1,2}_{\mathrm{loc}}(f(\Omega), \mathbb{R}^2)$, see [3], [7]. In [15], it was further shown that the regularity assumption $f \in W^{1,2}_{\mathrm{loc}}(\Omega, \mathbb{R}^2)$ can be slightly relaxed, say, to $|Df|^2 \log^{-1}(e + |Df|) \in L^1_{\mathrm{loc}}$. These results were based on a duality argument, relying on integration by parts against the Jacobian determinant J_f, that does not work when one only assumes that $f \in W^{1,p}_{\mathrm{loc}}(\Omega, \mathbb{R}^2)$ for some $p < 2$. The proofs of the first three theorems above are thus based on a different ingredient. The real problem is to prove that the distributional derivatives of the inverse mapping are indeed functions. This is obtained through delicate change of variable arguments; notice that in the setting of Theorem 1.2 f may well map a set of zero area to a set of positive area and a set of positive area to a set of zero area. If f preserves the class of sets of area zero, the situation is naturally substantially easier [19].

Recall that the statement of Theorem 1.2 is symmetric. Thus one could inquire if some power of the distortion of f^{-1} in Theorem 1.4 is also integrable. Our next result from [9], [6] shows that such a conclusion holds under exponen-

tial integrability of K_f, and the arguments in [9] show that one indeed needs exponential integrability.

Theorem 1.6. *Let $\Omega \subset \mathbb{R}^2$ be a domain and let $f \in W^{1,1}_{\rm loc}(\Omega, \mathbb{R}^2)$ be a homeomorphism of finite distortion. Assume that the distortion function K_f satisfies $\exp(\lambda K_f) \in L^1_{\rm loc}(\Omega)$, for some $\lambda > 0$. Then $K^p_{f^{-1}} \in L^1_{\rm loc}(f(\Omega))$ for all $p < \lambda$. Moreover, the claim may fail when $p = \lambda$.*

Let us close this section with some comments on the zero set of the Jacobian of our homeomorphism f. First of all, given any $0 < p < 1$, one can construct a homeomorphism f of finite distortion K_f so that K^p_f is locally integrable and the Jacobian of f vanishes on a set of positive area (then, necessarily, f maps a set of positive area onto a set of area zero). On the other hand, if K_f is locally integrable, then the Jacobian of f cannot vanish on a set of positive area. In fact, given $p \geq 1$, local integrability of K^p_f guarantees the local integrability of $\log^p(e + 1/J_f)$. For all this, see [14], [16]. What then about the possibility of the Jacobian being zero almost everywhere? This cannot happen for a homeomorphism of finite distortion because then also the partial derivatives would have to be zero almost everywhere, which would force our mapping to be a constant mapping. However, for every $1 \leq p < 2$ there exist homeomorphisms $f \in W^{1,p}_{\rm loc}(\Omega, \mathbb{R}^2)$ whose Jacobians equal zero almost everywhere, see [8], but no such homeomorphism can exist when $p \geq 2$.

2. Planar BV-mappings

Recall the Lipschitz homeomorphism $f(x_1, x_2) = (h(x_1), x_2)$ from the previous section. As discussed earlier, f^{-1} fails to be of the class $W^{1,1}_{\rm loc}$. However, it is easy to check that f^{-1} is of locally bounded variation, $f^{-1} \in BV_{\rm loc}(f(\Omega), \mathbb{R}^2)$.

Let us recall the definition of a mapping of locally bounded variation. Given a domain G and a mapping $g : G \to \mathbb{R}^2$, we say that g has bounded variation, $g \in BV(G, \mathbb{R}^2)$, if both component functions of g belong to the space $BV(G)$. This means that the distributional partial derivatives of each component function h of g are measures with finite total variation in G : there are Radon (signed) measures μ_1, μ_2 defined in G so that for $i = 1, 2$ $|\mu_i|(G) < \infty$ and

$$\int_G h D_i \varphi \, dx = - \int_G \varphi \, d\mu_i$$

for all $\varphi \in C^\infty_0(G)$. The gradient of h is then a vector-valued measure with finite total variation

$$||Dh|| = \sup\Big\{ \int_G h \, \mathrm{div}\, v \, dx : \ v = (v_1, v_2) \in C^\infty_0(G, \mathbb{R}^2),$$

$$|v(x)| \leq 1 \text{ for } x \in G \Big\} < \infty.$$

The total variation of $||Dh||$ can be considered as a Radon measure: given $A \subset G$ we set

$$||Dh||(A) = \sup\Big\{ \int_G h \operatorname{div} v \; dx : \; v = (v_1, v_2) \in C_0^\infty(G, \mathbb{R}^2),$$

$$|v(x)| \leq \chi_A(x) \text{ for } x \in \Omega \Big\}.$$

If $h \in W^{1,1}(G)$, then $||Dh||(A) = \int_A |\nabla u| \; dx$. For all this see [1]. Further, $g \in BV_{\rm loc}(G, \mathbb{R}^2)$ requires that $f \in BV(G', \mathbb{R}^2)$ for each subdomain $G' \subset\subset G$.

The following result from [12] can be viewed to be an analog of Theorem 1.2.

Theorem 2.1. *Let $\Omega, \; \Omega' \subset \mathbb{R}^2$ be domains and suppose that $f : \Omega \to \Omega'$ is a homeomorphism. Then $f \in BV_{\rm loc}(\Omega, \mathbb{R}^2)$ if and only if $f^{-1} \in BV_{\rm loc}(\Omega', \mathbb{R}^2)$. Moreover, both f and f^{-1} are differentiable almost everywhere.*

Since each homeomorphism $f \in W^{1,1}_{\rm loc}(\Omega, \mathbb{R}^2)$ belongs to $BV_{\rm loc}(\Omega, \mathbb{R}^2)$, we conclude from the above theorem, that the inverse of each planar Sobolev-homeomorphism is of locally bounded variation (but not necessarily of Sobolev-class $W^{1,1}_{\rm loc}$).

3. Mappings in Higher Dimensions

It should come as no big surprise that the results from the previous sections do not extend as such to higher dimensions. Indeed, given p strictly less $n-1$, where $n \geq 3$ is the dimension of our Euclidean space, it is not hard to construct homeomorphisms with p-integrable distributional derivatives so that the inverse mappings are not of locally bounded variation. For this see [12]. As observerved in [10], [12], [18], slightly stronger regularity assumptions are sufficient, but the following optimal result [4] was not proven until very recently.

Theorem 3.1. *Let $\Omega \subset \mathbb{R}^n$ be a domain and suppose that $f \in W^{1,n-1}_{\rm loc}(\Omega, \mathbb{R}^n)$ is a homeomorphism. Then $f^{-1} \in BV_{\rm loc}(f(\Omega), \mathbb{R}^n)$. If f furthermore has finite distortion, then $f^{-1} \in W^{1,1}_{\rm loc}(f(\Omega), \mathbb{R}^n)$ and has finite distortion.*

Theorem 1.2 was a crucial tool for the further results, except for Theorem 1.5 and Theorem 1.6, discussed in Section 1. Similarly, Theorem 3.1 allows one to prove higher dimensional versions of those results, see for example [10], [4].

References

[1] L. Ambrosio, N. Fusco and D. Pallara, *Functions of bounded variation and free discontinuity problems,* Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.

[2] K. Astala, T. Iwaniec, G.J. Martin and J. Onninen, *Extremal mappings of finite distortion,* Proc. London Math. Soc. (3) 91 (2005), no. 3, 655–702.

[3] B. Bojarski and T. Iwaniec, *Analytical foundations of the theory of quasiconformal mappings in R^n,* Ann. Acad. Sci. Fenn. Ser. A I Math. 8 (1983), no. 2, 257–324.

[4] M. Csörnyei, S. Hencl and J. Malý, *Homeomorphisms in the Sobolev space $W^{1,n-1}$,* to appear in J. Reine Angew. Math.

[5] F.W. Gehring and O. Lehto, *On the total differentiability of functions of a complex variable,* Ann. Acad. Sci. Fenn. A. I. Math 272 (1959), 1–9.

[6] J. Gill, *Integrability of derivatives of inverses of maps of exponentially integrable distortion in the plane,* J. Math. Anal. Appl. 352 (2009), 762–766.

[7] J. Heinonen and P. Koskela, *Sobolev mappings with integrable dilatations,* Arch. Rational Mech. Anal. 125 (1993), no. 1, 81–97.

[8] S. Hencl, *Sobolev homeomorphism with zero Jacobian almost everywhere,* to appear.

[9] S. Hencl and P. Koskela, *Regularity of the inverse of a planar Sobolev homeomorphism,* Arch. Ration. Mech. Anal. 180 (2006), no. 1, 75–95.

[10] S. Hencl, P. Koskela and J. Malý, *Regularity of the inverse of a Sobolev homeomorphism in space,* Proc. Roy. Soc. Edinburgh Sect. A 136 (2006), no. 6, 1267–1285.

[11] S. Hencl, P. Koskela and J. Onninen, *A note on extremal mappings of finite distortion,* Math. Res. Lett. 12 (2005), no. 2–3, 231–237.

[12] S. Hencl, P. Koskela and J. Onninen, *Homeomorphisms of bounded variation,* Arch. Ration. Mech. Anal. 186 (2007), no. 3, 351–360.

[13] T. Iwaniec and G. Martin, *Geometric Function Theory and Non-linear Analysis,* Oxford Science Publications, Clarendon Press, Oxford, 2001.

[14] P. Koskela and J. Malý, *Mappings of finite distortion: the zero set of the Jacobian,* J. Eur. Math. Soc., 5 (2003) no. 2, 95–105.

[15] P. Koskela and J. Onninen, *Mappings of finite distortion: capacity and modulus inequalities,* J. Reine Angew. Math. 599 (2006), 1–26.

[16] P. Koskela and J. Onninen, *Mappings of finite distortion: decay of the Jacobian in the plane,* Adv. Calc. Var. 1 (2008), no. 3, 309–321.

[17] D. Menchoff, *Sur les differentielles totales des fonctiones univalentes,* Math. Ann. 105 (1931), 75–85.

[18] J. Onninen, *Regularity of the inverse of spatial mappings with finite distortion,* Calc. Var. Partial Differential Equations 26 (2006), no. 3, 331–341.

[19] W.P. Ziemer, *Change of variables for absolutely continuous functions,* Duke Math. J. 36 1969 171–178.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Multiple Orthogonal Polynomials in Random Matrix Theory

Arno B.J. Kuijlaars*

Abstract

Multiple orthogonal polynomials are a generalization of orthogonal polynomials in which the orthogonality is distributed among a number of orthogonality weights. They appear in random matrix theory in the form of special determinantal point processes that are called multiple orthogonal polynomial (MOP) ensembles. The correlation kernel in such an ensemble is expressed in terms of the solution of a Riemann-Hilbert problem, that is of size $(r+1) \times (r+1)$ in the case of r weights.

A number of models give rise to a MOP ensemble, and we discuss recent results on models of non-intersecting Brownian motions, Hermitian random matrices with external source, and the two matrix model. A novel feature in the asymptotic analysis of the latter two models is a vector equilibrium problem for two or more measures, that describes the limiting mean eigenvalue density. The vector equilibrium problems involve both an external field and an upper constraint.

Mathematics Subject Classification (2010). Primary 42C05; Secondary 15B52, 31A15, 60C05, 60G55.

Keywords. Multiple orthogonal polynomials, non-intersecting Brownian motion, random matrices with external source, two matrix model, vector equilibrium problems, Riemann-Hilbert problem, steepest descent analysis.

1. Introduction

1.1. Random matrix theory. The Gaussian Unitary Ensemble (GUE) is the most prominent and most studied ensemble in random matrix theory. It is

*The author was supported in part by FWO-Flanders project G.0427.09, by K.U. Leuven research grant OT/08/33, by the Belgian Interuniversity Attraction Pole P06/02, and by grant MTM2008-06689-C02-01 of the Spanish Ministry of Science and Innovation.

Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200 B, 3001 Leuven, Belgium. E-mail: arno.kuijlaars@wis.kuleuven.be.

a probability measure on $n \times n$ Hermitian matrices for which the joint eigenvalue probability density function (p.d.f.) has the explicit form

$$\frac{1}{Z_n} \prod_{1 \le j < k \le n} (x_k - x_j)^2 \prod_{j=1}^{n} e^{-\frac{n}{2} x_j^2} \tag{1}$$

where Z_n is an explicitly known constant. The density (1) can be analyzed with the help of Hermite polynomials. Due to this connection with classical orthogonal polynomials many explicit calculations can be done, both for finite n and in the limit $n \to \infty$, see [40]. In particular it leads to a description of the limiting behavior of eigenvalues on the global (macroscopic) scale as well as on the local (microscopic) scale. The global scale is given by the well-known Wigner semi-circle law

$$\rho(x) = \frac{1}{2\pi} \sqrt{4 - x^2} \qquad -2 \le x \le 2, \tag{2}$$

in the sense that for eigenvalues $x_1, \ldots, x_n$ taken from (1), the empirical eigenvalue distribution $\frac{1}{n} \sum_{j=1}^{n} \delta(x_j)$ converges weakly to $\rho(x)$ almost surely as $n \to \infty$.

The local scale is characterized by the sine kernel

$$\mathcal{S}(x, y) = \frac{\sin \pi(x - y)}{\pi(x - y)} \tag{3}$$

in the bulk. This means that for any given $x^* \in (-2, 2)$ and any fixed $m \in \mathbb{N}$, the m-point correlation function (i.e., the marginal distribution)

$$R_{m,n}(x_1, \ldots, x_m) = \frac{n!}{(n-m)!} \int_{\mathbb{R}^{n-m}} \left[\frac{1}{Z_n} \prod_{1 \le j < k \le n} (x_k - x_j)^2 \prod_{j=1}^{n} e^{-\frac{1}{2} n x_j^2} \right] dx_{m+1} \cdots dx_n \tag{4}$$

has the scaling limit

$$\lim_{n \to \infty} \frac{1}{[\rho(x^*)n]^m} R_{m,n} \left(x^* + \frac{x_1}{\rho(x^*)n}, \ldots, x^* + \frac{x_m}{\rho(x^*)n} \right) = \det \left[\mathcal{S}(x_i, x_j) \right]_{1 \le i,j \le m}. \tag{5}$$

At the edge points ± 2 the sine kernel (3) is replaced by the Airy kernel

$$\mathcal{A}(x, y) = \frac{\operatorname{Ai}(x) \operatorname{Ai}'(y) - \operatorname{Ai}'(x) \operatorname{Ai}(y)}{x - y} \tag{6}$$

and a scaling limit as in (5) (with scaling factor $cn^{2/3}$ instead of $\rho(x^*)n$) holds for $x^* = \pm 2$. This result leads in particular to the statement about the largest eigenvalue

$$\lim_{n \to \infty} \operatorname{Prob} \left(\max_{1 \le k \le n} x_k < 2 + \frac{t}{cn^{2/3}} \right) = \det \left[I - \mathcal{A}_{(t, \infty)} \right] \tag{7}$$

where $\mathcal{A}$ is the Airy kernel (6) and the determinant is the Fredholm determinant of the integral operator with Airy kernel acting on $L^2(t,\infty)$. The limiting distribution (7) is the famous Tracy-Widom distribution named after the authors of the seminal work [45] in which the right-hand side of (7) is expressed in terms of the Hastings-McLeod solution of the Painlevé II equation.

These basic results of random matrix theory have been extended and generalized in numerous directions. Within the theory of random matrices, they have been generalized to ensembles with unitary, orthogonal and symplectic symmetry and to non-invariant ensembles (Wigner ensembles). The distribution functions of random matrix theory also appear in many other probabilistic models that have no apparent connection with random matrices (models of non-intersecting paths, tiling models, and stochastic growth models), see e.g. [8], [31].

Mehta's book [40] is the standard reference on random matrix theory. The book of Deift [22] has been very influential in introducing Riemann-Hilbert techniques into the study of random matrices. In recent years, a number of new monographs appeared [2], [6], [16], [23], [29] that cover the various aspects of the theory of random matrices.

1.2. Unitary ensembles and orthogonal polynomials. One direction within random matrix theory is the study of ensembles of the form

$$\frac{1}{Z_n} e^{-n \operatorname{Tr} V(M)}\, dM \tag{8}$$

defined on $n \times n$ Hermitian matrices M, which reduces to the GUE in case $V(x) = \frac{1}{2}x^2$. The ensembles (8) have the property of unitary invariance and are called unitary ensembles. The eigenvalues have the p.d.f.

$$\frac{1}{Z_n} \prod_{1\le j<k\le n} (x_k - x_j)^2 \prod_{j=1}^{n} e^{-nV(x_j)} \tag{9}$$

with a different normalizing constant Z_n. [Throughout, we use Z_n to denote a normalizing constant, which may be different from one formula to the next.]

Again explicit calculations can be done due to the connection with orthogonal polynomials [23], [40]. For a given n, we consider the monic polynomial $P_{k,n}$ of degree k that satisfies

$$\int_{-\infty}^{\infty} P_{k,n}(x) x^j e^{-nV(x)} dx = h_{k,n}\delta_{j,k}, \qquad j = 0, \dots, k.$$

Then (9) is a determinantal point process [2], [44], with kernel

$$K_n(x,y) = \sqrt{e^{-nV(x)}}\sqrt{e^{-nV(y)}} \sum_{k=0}^{n-1} \frac{P_{k,n}(x)P_{k,n}(y)}{h_{k,n}} \tag{10}$$

which means that for every $m \in \mathbb{N}$ the m-point correlation functions, defined as in (4), have the determinantal form

$$\det \left[K_n(x_i, x_j)\right]_{i,j=1,\ldots,m}.$$

As $n \to \infty$, the limiting mean eigenvalue density

$$\rho(x) = \lim_{n\to\infty} \frac{1}{n} K_n(x,x)$$

is no longer Wigner's semi-circle law (2), but instead it is the density ρ of the probability measure μ that minimizes the weighted logarithmic energy

$$\iint \log \frac{1}{|x-y|} d\mu(x) d\mu(y) + \int V(x) d\mu(x) \tag{11}$$

among all probability measures on $\mathbb{R}$.

Local eigenvalue statistics, however, have a universal behavior as $n \to \infty$, that is described by the sine kernel (3) in the bulk. Thus for points x^* with $\rho(x^*) > 0$ the limit (5) holds true. At edge points of the limiting spectrum the density ρ typically vanishes as a square root and then the universal Airy kernel (6) appears. For real analytic potentials V this was proved in [11], [24] using Riemann-Hilbert methods. This was vastly extended to non-analytic potentials in recent works of Lubinsky [38] and Levin and Lubinsky [37], among many others.

1.3. This paper. In this paper we present an overview of the work (mainly of the author and co-workers) on multiple orthogonal polynomials and their relation to random matrix theory. Multiple orthogonal polynomials are a generalization of orthogonal polynomials that have their origins in approximation theory (Hermite-Padé approximation), see e.g. [3, 42].

They enter the theory of random matrices via a generalization of (9) which we call a multiple orthogonal polynomial (MOP) ensemble [34]. We present a number of models that give rise to a MOP ensemble, namely the model of non-intersecting Brownian motions, the random matrix model with external source and the two matrix model.

The MOPs are described by a Riemann-Hilbert problem that may be used for asymptotic analysis as $n \to \infty$ by extending the Deift-Zhou method of steepest descent [25]. The extensions are non-trivial and involve either an a priori knowledge of an underlying Riemann surface (the spectral curve) or the formulation of a relevant equilibrium problem from logarithmic potential theory [43], which asks for a generalization of the weighted energy functional (11).

The latter approach has been succesfully applied to the random matrix model with external source and to the two matrix model, but only in very special cases, as will be discussed at the end of the paper.

2. Multiple Orthogonal Polynomials

2.1. MOP ensemble. We will describe here multiple orthogonal polynomials of type II, which we simply call multiple orthogonal polynomials. There is also a dual notion of type I multiple orthogonal polynomials.

Suppose we have a finite number of weight functions $w_1, \ldots, w_r$ on $\mathbb{R}$ and a multi-index $\vec{n} = (n_1, \ldots, n_r) \in \mathbb{N}^r$. Associated with these data is the monic polynomial $P_{\vec{n}}$ of degree $|\vec{n}| = n_1 + \cdots + n_r$ so that

$$\int_{-\infty}^{\infty} P_{\vec{n}}(x) x^j w_k(x)\, dx = 0, \quad \text{for } j = 0, \ldots, n_k - 1, \quad k = 1, \ldots, r. \tag{12}$$

The linear system of equations (12) may not be always uniquely solvable, but in many important cases it is. If $P_{\vec{n}}$ uniquely exists then it is called the multiple orthogonal polynomial (MOP) associated with the weights $w_1, \ldots, w_r$ and multi-index $\vec{n}$.

Existence and uniqueness does hold in the following situation. Assume that

$$\frac{1}{Z_n} \det\left[f_j(x_k)\right]_{j,k=1,\ldots,n} \left[\prod_{1 \le j < k \le n} (x_k - x_j)\right] \tag{13}$$

is a p.d.f. on $\mathbb{R}^n$, where $n = |\vec{n}|$ and the linear span of the functions $f_1, \ldots, f_n$ is the same as the linear span of the set of functions

$$\{x^j w_k(x) \mid j = 0, \ldots, n_k - 1,\ k = 1, \ldots, r\}.$$

So the assumption is that (13) is non-negative for every choice of $x_1, \ldots, x_n \in \mathbb{R}^n$, and that the normalization constant Z_n can be taken so that the integral (13) over $\mathbb{R}^n$ is equal to one. Then the MOP satisfying (12) exists and is given by

$$P_{\vec{n}}(x) = \mathbb{E}\left[\prod_{j=1}^{n} (x - x_j)\right].$$

We call a p.d.f. on $\mathbb{R}^n$ of the form (13) a MOP ensemble, see [34].

2.2. Correlation kernel and RH problem. The MOP ensemble (13) is a determinantal point process [44] (more precisely a biorthogonal ensemble [17]) with a correlation kernel K_n that is constructed out of multiple orthogonal polynomials of type II and type I. It is conveniently described in terms of the solution of a Riemann-Hilbert (RH) problem. This RH problem for MOPs [47] is a generalization of the RH problem for orthogonal polynomials due to Fokas, Its, and Kitaev [28].

The RH problem asks for an $(r+1)\times(r+1)$ matrix valued function Y so that

$$\begin{cases} \bullet\ Y:\mathbb{C}\setminus\mathbb{R}\to\mathbb{C}^{(r+1)\times(r+1)} \text{ is analytic,} \\ \bullet\ Y \text{ has limiting values on } \mathbb{R}, \text{ denoted by } Y_+ \text{ and } Y_-, \text{ where } Y_\pm(x) \text{ is} \\ \quad \text{the limit of } Y(z) \text{ as } z\to x\in\mathbb{R} \text{ with } \pm\operatorname{Im} z>0, \text{ satisfying} \\ \qquad Y_+(x)=Y_-(x)\begin{pmatrix} 1 & w_1(x) & \cdots & w_r(x) \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \text{ for } x\in\mathbb{R}, \\ \bullet\ Y(z)=(I+O(1/z))\operatorname{diag}\begin{pmatrix} z^n & z^{-n_1} & \cdots & z^{-n_r}\end{pmatrix} \text{ as } z\to\infty. \end{cases} \tag{14}$$

If the MOP $P_{\vec{n}}$ with weights $w_1,\dots,w_r$ and multi-index $\vec{n}=(n_1,\dots,n_r)$ exists then the RH problem (14) has a unique solution. If the MOPs with multi-indices $\vec{n}-\vec{e}_j$ also exist, where $\vec{e}_j$ is the jth unit vector of length r, then the first column of Y consists of

$$Y_{1,1}(z)=P_{\vec{n}}(z), \qquad Y_{j+1,1}(z)=c_{j,\vec{n}}P_{\vec{n}-\vec{e}_j}(z), \qquad j=1,\dots,r \tag{15}$$

where $c_{j,\vec{n}}$ is the constant

$$c_{j,\vec{n}}=-2\pi i\left[\int_{-\infty}^{\infty}P_{\vec{n}-\vec{e}_j}(x)x^{n_j-1}w_j(x)dx\right]^{-1}\neq 0.$$

The other columns of Y contain Cauchy transforms

$$Y_{j,k+1}(z)=\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{Y_{j,1}(x)w_k(x)}{x-z}dx, \qquad j=1,\dots,r+1, \quad k=1,\dots,r.$$

It is a remarkable fact that the correlation kernel of the MOP ensemble (13) is expressed as follows in terms of the solution of the RH problem, see [20],

$$K_n(x,y)=\frac{1}{2\pi i(x-y)}\begin{pmatrix}0 & w_1(y) & \cdots & w_r(y)\end{pmatrix}Y_+^{-1}(y)Y_+(x)\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}, \quad x,y\in\mathbb{R}. \tag{16}$$

The inverse matrix Y^{-1} contains MOPs of type I, and the formula (16) is essentially the Christoffel-Darboux formula for multiple orthogonal polynomials.

Besides giving a concise formula for the correlation kernel, the expression (16) for the kernel gives also a possible way to do asymptotic analysis in view of the Deift-Zhou method of steepest descent for RH problems.

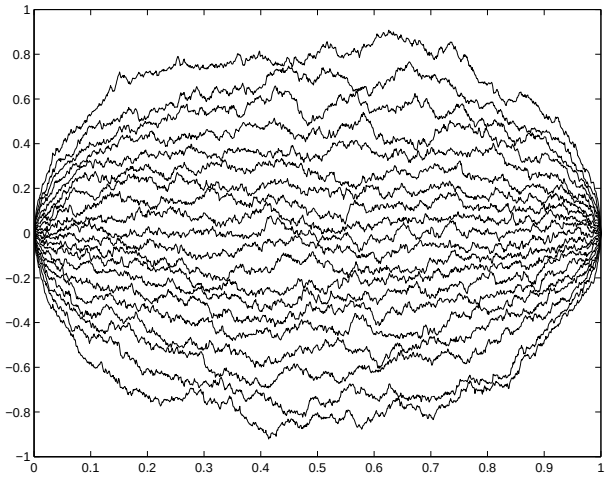


Figure 1. Non-intersecting Brownian bridges starting and ending at 0. At any intermediate time $t \in (0,1)$ the positions of the paths have the same distribution as the (appropriately rescaled) eigenvalues of an $n \times n$ GUE matrix.

3. Non Intersecting Path Ensembles

A rich source of examples of determinantal point processes is provided by non-intersecting path ensembles. In special cases these reduce to MOP ensembles.

3.1. Non-intersecting Brownian motion. Consider a one-dimensional strong Markov process with transition probability densities $p_t(x,y)$ for $t > 0$. Suppose n independent copies are given with respective starting values $a_1 < a_2 < \cdots < a_n$ at time $t = 0$ and prescribed ending values $b_1 < b_2 < \cdots < b_n$ at time $t = T > 0$ that are conditioned not to intersect in the full time interval $0 < t < T$. Then by an application of a theorem of Karlin and McGregor [32], the positions of the paths at an intermediate time $t \in (0,T)$ have the joint p.d.f.

$$\frac{1}{Z_n} \det\left[p_t(a_j, x_k)\right]_{1\le j,k\le n} \cdot \det\left[p_{T-t}(x_k, b_l)\right]_{1\le k,l\le n}. \tag{17}$$

In a discrete combinatorial setting the result of Karlin and McGregor is known as the Lindstrom-Gessel-Viennot theorem.

The density function (17) is a biorthogonal ensemble, which in very special cases reduces to the form (13) of a MOP ensemble.

An example is the case of Brownian motion (actually Brownian bridges) with the transition probability density

$$p_t(x,y) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}}, \qquad t > 0.$$

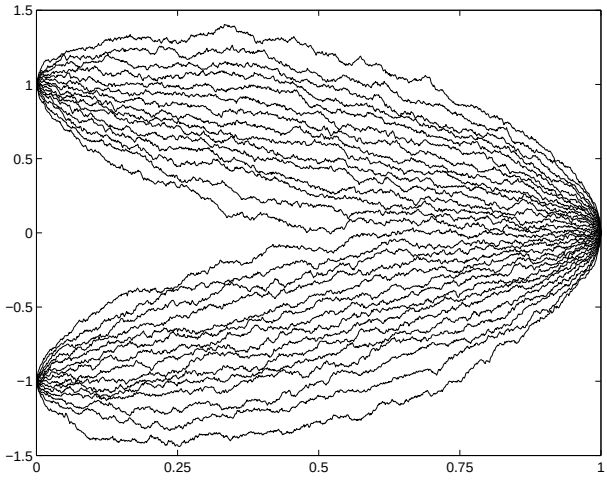


Figure 2. Non-intersecting Brownian bridges starting at two different values and ending at 0. At any time $t \in (0,1)$ the positions of the paths have the same distribution as the eigenvalues of an $n \times n$ GUE matrix with external source. The distribution is a multiple Hermite ensemble with two Gaussian weights (18).

In the confluent limit where all $a_j \to 0$ and all $b_l \to 0$ the p.d.f. (17) turns into

$$\frac{1}{Z_n} \prod_{1 \le j < k \le n} (x_k - x_j)^2 \prod_{j=1}^{n} e^{-\frac{T}{2t(T-t)} x_j^2}$$

with a different constant Z_n. This is up to trivial scaling the same as the p.d.f. (1) for the eigenvalues of an $n \times n$ GUE matrix.

If however, we let all $b_l \to 0$ and choose only r different starting values, denoted by $a_1, \ldots, a_r$, and n_j paths start at a_j, then (17) turns into a MOP ensemble with weights

$$w_j(x) = e^{-\frac{T}{2t(T-t)} x^2 + \frac{a_j}{t} x}, \qquad j = 1, \ldots, r, \tag{18}$$

and multi-index $(n_1, \ldots, n_r)$. This is a multiple Hermite ensemble, since the associated MOPs are multiple Hermite polynomials [5]

3.2. Non-intersecting squared Bessel paths. The squared Bessel process is another one-dimensional Markov process which gives rise to a MOP ensemble. The squared Bessel process is a Markov process on $[0, \infty)$, depending on a parameter $\alpha > -1$, with transition probability density

$$p_t(x,y) = \frac{1}{2t} \left(\frac{y}{x}\right)^{\alpha/2} e^{-\frac{1}{2t}(x+y)} I_\alpha \left(\frac{\sqrt{xy}}{t}\right), \qquad x, y > 0,$$

where I_α is the modified Bessel function of first kind of order α. In the limit where all $a_j \to a > 0$ and $b_j \to 0$ the p.d.f. (17) for the positions of the paths at time $t \in (0,T)$ is a MOP ensemble with two weights

$$w_1(x) = x^{\alpha/2} e^{-\frac{T}{2t(T-t)}x} I_\alpha\left(\frac{\sqrt{ax}}{t}\right)$$
$$w_2(x) = x^{(\alpha+1)/2} e^{-\frac{T}{2t(T-t)}x} I_{\alpha+1}\left(\frac{\sqrt{ax}}{t}\right)$$

and multi-index (n_1, n_2) where $n_1 = \lceil n/2 \rceil$ and $n_2 = \lfloor n/2 \rfloor$, see [35]. In the limit $a \to 0$ this further reduces to an orthogonal polynomial ensemble for a Laguerre weight.

4. Random Matrix Models

The random matrix model with external source, and the two matrix model also give rise to MOP ensembles.

4.1. Random matrices with external source. The Hermitian matrix model with external source is the probability measure

$$\frac{1}{Z_n} e^{-n \operatorname{Tr}(V(M) - AM)} dM \tag{19}$$

on $n \times n$ Hermitian matrices, where the external source A is a given Hermitian $n \times n$ matrix. This is a modification of the usual Hermitian matrix model, in which the unitary invariance is broken [18], [48].

Due to the Harish-Chandra/Itzykson-Zuber integral formula [30], it is possible to integrate out the eigenvectors explicitly. In case the eigenvalues $a_1, \ldots, a_n$ of A are all distinct, we obtain the explicit p.d.f.

$$\frac{1}{Z_n} \det\left[e^{n a_i x_j}\right]_{1 \le i,j \le n} \cdot \prod_{1 \le j < k \le n} (x_k - x_j) \cdot \prod_{j=1}^{n} e^{-nV(x_j)}$$

for the eigenvalues of M. In case that $a_1, \ldots, a_r$ are the distinct eigenvalues of A, with respective multiplicities $n_1, \ldots, n_r$, then the eigenvalues of M are distributed as a MOP ensemble (13) with weights

$$w_j(x) = e^{-n(V(x) - a_j x)}, \qquad j = 1, \ldots, r \tag{20}$$

and multi-index $(n_1, \ldots, n_r)$, see [13]

For the case $V(x) = \frac{1}{2}x^2$ the external source model (19) is equivalent to the model of non-intersecting Brownian motions with several starting points and one ending point, cf. (18).

4.2. Two matrix model. The Hermitian two matrix model

$$\frac{1}{Z_n} e^{-n \operatorname{Tr}(V(M_1)+W(M_2)-\tau M_1 M_2)}\, dM_1 dM_2 \tag{21}$$

is a probability measure defined on couples (M_1, M_2) of $n \times n$ Hermitian matrices. Here V and W are two potentials (typically polynomials) and $\tau \neq 0$ is a coupling constant. The model is of great interest in 2d quantum gravity [21], [30], [33], as it allows for a large class of critical phenomena.

The eigenvalues of the matrices M_1 and M_2 are fully described by biorthogonal polynomials. These are two sequences $(P_{k,n})_k$ and $(Q_{j,n})_j$ of monic polynomials, $\deg P_{k,n} = k$, $\deg Q_{j,n} = j$, such that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{k,n}(x) Q_{j,n}(y) e^{-n(V(x)+W(y)-\tau xy)}\, dx\, dy = 0, \qquad \text{if } j \neq k, \tag{22}$$

see e.g. [9], [27], [40], [41].

If W is a polynomial then the biorthogonality conditions (22) can be seen as multiple orthogonal polynomial conditions with respect to $r = \deg W - 1$ weights

$$w_{j,n}(x) = e^{-nV(x)} \int_{-\infty}^{\infty} y^j e^{-n(W(y)-\tau xy)} dy, \qquad j = 0, \ldots, r-1, \tag{23}$$

see [36]. Furthermore, the eigenvalues of M_1 are a MOP ensemble (13) with the weights (23) and multi-index $\vec{n} = (n_0, \ldots, n_{r-1})$ with $n_j = \lceil n/r \rceil$ for $j = 0, \ldots, q-1$ and $n_j = \lfloor n/r \rfloor$ for $j = q, \ldots, r-1$ if $n = pr + q$ with p and $0 \leq q < r$ non-negative integers, see [26] for the case where $W(y) = \frac{y^4}{4}$.

5. Large n Behavior and Critical Phenomena

We discuss the large n behavior in the above described models.

5.1. Non-intersecting Brownian motion. In order to have interesting limit behavior as $n \to \infty$ in the non-intersecting Brownian motion model we scale the time variables $T \mapsto 1/n$, $t \mapsto t/n$, so that $0 < t < 1$. In the case of one starting value and one ending value, see Figure 1, the paths will fill out an ellipse as $n \to \infty$.

In the situation of Figure 2 the paths fill out a heart-shaped region as $n \to \infty$, as shown in Figure 3. New critical behavior appears at the cusp point where the two groups of paths come together and merge into one.

Around the critical time the correlation kernels have a double scaling limit, which is given by the one-parameter family of Pearcey kernels

$$\mathcal{P}(x, y; b) = \frac{p(x)q''(y) - p'(x)q'(y) + p''(x)q(y) - bp(x)q(y)}{x - y}, \qquad b \in \mathbb{R}, \tag{24}$$

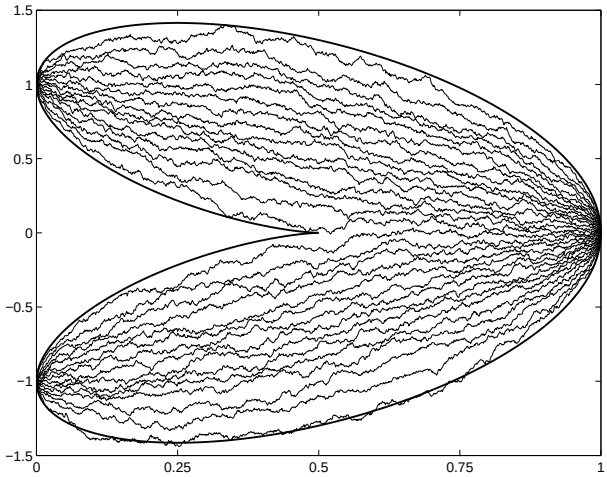


Figure 3. Non-intersecting Brownian bridges starting at two different values and ending at 0. As $n \to \infty$, the paths fill out a heart-shaped domain. Critical behavior at the cusp point is desribed by the Pearcey kernel (24).

where p and q are solutions of the Pearcey equations $p'''(x) = xp(x) - bp'(x)$ and $q'''(y) = yq(y) + bq'(y)$. This kernel was first identified by Brézin and Hikami [18] who also gave the double integral representation

$$\mathcal{P}(x,y) = \frac{1}{(2\pi i)^2} \int_C \int_{-i\infty}^{i\infty} e^{-\frac{1}{4}s^4 + \frac{b}{2}s^2 - ys + \frac{1}{4}t^4 - \frac{b}{2}t^2 + xt} \frac{ds\,dt}{s-t} \tag{25}$$

where the contour C consists of the rays from $\pm\infty e^{i\pi/4}$ to 0 and the rays from 0 to $\pm\infty e^{-i\pi/4}$.

Consideration of multiple times near the critical time leads to an extended Pearcey kernel and the Pearcey process given by Tracy and Widom [46].

As already noted above, the model of non-intersecting Brownian motion with two starting points and one ending point is related to the Gaussian random matrix model with external source

$$\frac{1}{Z_n} e^{-n \operatorname{Tr}(\frac{1}{2}M^2 - AM)}\, dM, \tag{26}$$

with external source

$$A = \operatorname{diag}(\underbrace{a, \ldots, a}_{n/2 \text{ times}}, \underbrace{-a, \ldots, -a}_{n/2 \text{times}}). \tag{27}$$

In this setting the critical a-value is $a_{crit} = 1$ and the Pearcey kernel (24) arises as $n \to \infty$ with $a = 1 + \frac{b}{2\sqrt{n}}$. In [15] this was studied with the use of the Riemann-Hilbert problem (14) for multiple Hermite polynomials with two

weights $e^{-n(\frac{1}{2}x^2 \pm ax)}$. The asymptotic analysis as $n \to \infty$ was done with an extension of the Deift-Zhou method of steepest descent [25] to the case of a 3×3 matrix valued RH problem. See also [14] and [4] for a steepest descent analysis of the RH problem in the non-critical regimes $a > 1$ and $0 < a < 1$, respectively.

Another interesting asymptotic regime is the model of non-intersecting Brownian motion with outliers. In this model a rational modification of the Airy kernel appears that was first described in [7] in the context of complex sample covariance matrices, see also [1].

5.2. Random matrices with external source. If V is quadratic in the random matrix model with external source (19) then this model can be mapped to the model of non-intersecting Brownian motions. Progress on this model beyond the quadratic case is due to McLaughlin [39] who found the spectral curve for the quartic potential $V(x) = \frac{1}{4}x^4$ and for a sufficiently large (again A is as in (27)).

A method based on a vector equilibrium problem was introduced recently by Bleher, Delvaux and Kuijlaars [10]. The vector equilibrium problem extends the equilibrium problem for the weighted energy (11) that is important for the unitary ensembles and which is crucial in the steepest descent analysis of the RH problem for orthogonal polynomials [24].

In [10] it is assumed that V is an even polynomial, and that A is again given as in (27). The vector equilibrium problem involves two measures μ_1 and μ_2, and it asks to minimize the energy functional

$$\iint \log \frac{1}{|x-y|} d\mu_1(x) d\mu_1(y) + \iint \log \frac{1}{|x-y|} d\mu_2(x) d\mu_2(y) - \iint \log \frac{1}{|x-y|} d\mu_1(x) d\mu_2(y) + \int (V(x) - a|x|)\, d\mu_1(x) \quad (28)$$

where μ_1 is on $\mathbb{R}$ with $\int d\mu_1 = 1$, μ_2 is on $i\mathbb{R}$ (the imaginary axis) with $\int d\mu_2 = 1/2$, and in addition $\mu_2 \leq \sigma$, where σ is the measure on $i\mathbb{R}$ with constant density

$$\frac{d\sigma}{|dz|} = \frac{a}{\pi}. \quad (29)$$

There is a unique minimizer, and the density ρ_1 of the measure μ_1 is the limiting mean eigenvalue density

$$\rho_1(x) = \lim_{n\to\infty} \frac{1}{n} K_n(x,x)$$

where K_n is the correlation kernel of the MOP ensemble with weights $e^{-n(V(x)\pm ax)}$. The RH problem (14) is analyzed in the large n limit with the Deift/Zhou steepest descent method in which the minimizers from the vector equilibrium problem play a crucial role.

The upper constraint $\mu_2 \leq \sigma$ is not active for large enough a and in that case the support of μ_1 has a gap around 0. For smaller values of a the constraint σ is active along an interval $[-ic, ic]$, $c > 0$, on the imaginary axis. Critical phenomena take place when either the constraint becomes active, or the gap around 0 closes, or both. If one of these two phenomena happens, then this generically will be a phase transition of the Painlevé II type that was described in the unitary matrix model in [12] and [19]. If the two phenomena happen simultaneously then this is expected to be phase transition of the Pearcey type which, if true, would be a confirmation of the universality of the Pearcey kernels (24) at the closing of a gap [18].

Both kinds of transitions are valid in the external source model with even quartic potential $V(x) = \frac{1}{4}x^4 - \frac{t}{2}x^2$, see [10]. For the particular value $t = \sqrt{3}$, there is a passage from the Painlevé II transition (for $t > \sqrt{3}$) to the Pearcey transition (for $t < \sqrt{3}$). The description of the phase transition for $t = \sqrt{3}$ remains open.

5.3. Two matrix model. In [26] Duits and Kuijlaars applied the steepest descent analysis to the RH problem for the two matrix model (21) with quartic potential

$$W(y) = \frac{1}{4}y^4 \tag{30}$$

and for V an even polynomial. The corresponding MOP ensemble has three weights of the form (23) and the RH problem (14) is of size 4×4. Again a vector equilibrium problem plays a crucial role.

The vector equilibrium problem in [26] involves three measures μ_1, μ_2 and μ_3. It asks to minimize the energy functional

$$\sum_{j=1}^{3} \iint \log \frac{1}{|x-y|} d\mu_j(x) d\mu_j(y) - \sum_{j=1}^{2} \iint \log \frac{1}{|x-y|} d\mu_j(x) d\mu_{j+1}(y) + \int (V(x) - \frac{3}{4}|\tau x|^{4/3})\, d\mu_1(x) \tag{31}$$

among measures μ_1 on $\mathbb{R}$ with $\int d\mu_1 = 1$, μ_2 on $i\mathbb{R}$ with $\int d\mu_2 = 2/3$ and μ_3 on $\mathbb{R}$ with $\int d\mu_3 = 1/3$. In addition $\mu_2 \leq \sigma$ where σ is a given measure on $i\mathbb{R}$ with density

$$\frac{d\sigma}{|dz|} = \frac{\sqrt{3}}{2\pi}|\tau|^{4/3}|z|^{1/3}, \qquad z \in i\mathbb{R}. \tag{32}$$

There is a unique minimizer and the density ρ_1 of the first measure μ_1 is equal to the limiting mean eigenvalue density of the matrix M_1 in the two matrix model. In addition, the usual scaling limits (sine kernel and Airy kernel) are valid in the local eigenvalue regime, see [26]. However there is no new critical behavior in the two matrix model with W is given by (30).

New multicritical behavior is predicted in [21] for more general potentials. For the more general quartic potential $W(y) = \frac{1}{4}y^4 - \frac{t}{2}y^2$ an approach based on a modification of the vector equilibrium problem (31) is under current investigation.

References

[1] M. Adler, J. Delépine, and P. van Moerbeke, Dyson's nonintersecting Brownian motions with a few outliers, Comm. Pure Appl. Math. 62 (2009), 334–395.

[2] G. Anderson, A. Guionnet, and O. Zeitouni, An Introduction to Random Matrices, Cambridge University Press, Cambridge, 2010.

[3] A.I. Aptekarev, Multiple orthogonal polynomials, J. Comput. Appl. Math. 99 (1998), 423–447.

[4] A.I. Aptekarev, P. M. Bleher and A.B.J. Kuijlaars, Large n limit of Gaussian random matrices with external source. II, Comm. Math. Phys. 259 (2005), 367–389.

[5] A.I. Aptekarev, A. Branquinho, and W. Van Assche, Multiple orthogonal polynomials for classical weights, Trans. Amer. Math. Soc. 355 (2003), 3887–3914.

[6] Z. Bai and J.W. Silverstein, Spectral analysis of large dimensional random matrices, 2nd ed., Springer Series in Statistics, Springer, New York, 2010.

[7] J. Baik, G. Ben Arous and S. Péché, Phase transition of the largest eigenvalue for non-null complex sample covariance matrices, Ann. Probab. 33 (2005), 1643–1697.

[8] J. Baik, P. Deift, and K. Johansson, On the distribution of the length of the longest increasing subsequence of random permutations, J. Amer. Math. Soc. 12 (1999), 1119–1178.

[9] M. Bertola, B. Eynard, and J. Harnad, Duality, biorthogonal polynomials and multi-matrix models, Comm. Math. Phys. 229 (2002), 73–120.

[10] P.M. Bleher, S. Delvaux, and A.B.J. Kuijlaars, Random matrix model with external source and a constrained vector equilibrium problem, preprint arXiv:1001.1238.

[11] P. Bleher and A. Its, Semiclassical asymptotics of orthogonal polynomials, Riemann-Hilbert problem, and universality in the matrix model, Ann. Math. 150 (1999), 185–266.

[12] P. Bleher and A. Its, Double scaling limit in the random matrix model: the Riemann-Hilbert approach, Comm. Pure Appl. Math. 56 (2003), 433–516.

[13] P.M. Bleher and A.B.J. Kuijlaars, Random matrices with external source and multiple orthogonal polynomials, Internat. Math. Research Notices 2004:3 (2004), 109–129.

[14] P.M. Bleher and A.B.J. Kuijlaars, Large n limit of Gaussian random matrices with external source I, Comm. Math. Phys. 252 (2004), 43–76.

[15] P.M. Bleher and A.B.J. Kuijlaars, Large n limit of Gaussian random matrices with external source III: double scaling limit, Comm. Math. Phys. 270 (2007), 481–517.

[16] G. Blower, Random Matrices: High Dimensional Phenomena, Cambridge University Press, Cambridge, 2009.

[17] A. Borodin, Biorthogonal ensembles, Nuclear Phys. B 536 (1999), 704–732.

[18] E. Brézin and S. Hikami, Universal singularity at the closure of a gap in a random matrix theory, Phys. Rev. E 57 (1998), 4140–4149.

[19] T. Claeys and A.B.J. Kuijlaars, Universality of the double scaling limit in random matrix models, Comm. Pure Appl. Math. 59 (2006), 1573–1603.

[20] E. Daems and A.B.J. Kuijlaars, A Christoffel-Darboux formula for multiple orthogonal polynomials, J. Approx. Theory 130 (2004), 188–200.

[21] J.-M. Daul, V.A. Kazakov, and I.K. Kostov, Rational theories of 2d gravity from the two-matrix model, Nuclear Phys. B 409 (1993), 311–338.

[22] P. Deift, Orthogonal Polynomials and Random Matrices: a Riemann-Hilbert approach, Courant Lecture Notes in Mathematics, Vol. 3, Amer. Math. Soc., Providence RI, 1999.

[23] P. Deift and D. Gioev, Random Matrix Theory: Invariant Ensembles and Universality, Courant Lecture Notes in Mathematics, Vol. 18, Amer. Math. Soc., Providence RI, 2009.

[24] P. Deift, T. Kriecherbauer, K.T-R. McLaughlin, S. Venakides, and X. Zhou, Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory, Comm. Pure Appl. Math. 52 (1999), 1335–1425.

[25] P. Deift and X. Zhou, A steepest descent method for oscillatory Riemann-Hilbert problems. Asymptotics for the MKdV equation, Ann. Math 137 (1993), 295–368.

[26] M. Duits and A.B.J. Kuijlaars, Universality in the two matrix model: a Riemann-Hilbert steepest descent analysis, Comm. Pure Appl. Math. 62 (2009), 1076–1153.

[27] N. Ercolani and K.T-R McLaughlin, Asymptotics and integrable structures for biorthogonal polynomials associated to a random two-matrix model, Physica D 152/153 (2001), 232–268.

[28] A.S. Fokas, A.R. Its, and A.V. Kitaev, The isomonodromy approach to matrix models in 2D quantum gravity, Commun. Math. Phys. 147 (1992), 395–430.

[29] P.J. Forrester, Log-Gases and Random Matrices, LMS Mongraphs Series Vol. 34, Princeton Univ. Press, Princeton N.J., 2010.

[30] C. Itzykson and J.B. Zuber, The planar approximation II, J. Math. Phys. 21 (1980), 411–421.

[31] K. Johansson, Discrete orthogonal polynomial ensembles and the Plancherel measure, Ann. Math. 153 (2001), 259–296.

[32] S. Karlin and J. McGregor, Coincidence probabilities, Pacific J. Math. 9 (1959), 1141–1164.

[33] V.A. Kazakov, Ising model on a dynamical planar random lattice: exact solution. Phys. Lett. A 119 (1986), 140–144.

[34] A.B.J. Kuijlaars, Multiple orthogonal polynomial ensembles, in "Recent Trends in Orthogonal Polynomials and Approximation Theory" (J. Arvesú et al., eds.), Contemp. Math. 507, 2010, pp. 155–171.

[35] A.B.J. Kuijlaars, A. Martínez-Finkelshtein and F. Wielonsky, Non-intersecting squared Bessel paths and multiple orthogonal polynomials for modified Bessel weights, Comm. Math. Phys. 286 (2009), 217–275.

[36] A.B.J. Kuijlaars and K.T-R McLaughlin, A Riemann-Hilbert problem for biorthogonal polynomials, J. Comput. Appl. Math. 178 (2005), 313–320.

[37] E. Levin and D.S. Lubinsky, Universality limits in the bulk for varying measures, Adv. Math. 219 (2008), 743–779.

[38] D.S. Lubinsky, A new approach to universality limits involving orthogonal polynomials, Ann. Math. 170 (2009), 915–939.

[39] K.T-R McLaughlin, Asymptotic analysis of random matrices with external source and a family of algebraic curves, Nonlinearity 20 (2007), 1547–1571.

[40] M.L. Mehta, Random Matrices, 3rd ed., Elsevier/Academic Press, Amsterdam, 2004.

[41] M.L. Mehta and P. Shukla, Two coupled matrices: eigenvalue correlations and spacing functions, J. Phys. A 27 (1994), 7793–7803.

[42] E.M. Nikishin and V.N. Sorokin, Rational Approximations and Orthogonality, Translations of Mathematical Monographs vol. 92, Amer. Math. Soc., Providence, RI, 1991.

[43] E.B. Saff and V. Totik, Logarithmic Potentials with External Fields, Grundlehren der Mathematischen Wissenschaften 136, Springer-Verlag, Berlin, 1997.

[44] A. Soshnikov, Determinantal random point fields, Russian Math. Surveys 55 (2000), 923–975.

[45] C. Tracy and H. Widom, Level-spacing distributions and the Airy kernel, Comm. Math. Phys. 159 (1994), 151–174.

[46] C. Tracy and H. Widom, The Pearcey process, Comm. Math. Phys. 263 (2006), 381–400.

[47] W. Van Assche, J. Geronimo, and A.B.J. Kuijlaars, Riemann-Hilbert problems for multiple orthogonal polynomials, in "Special Functions 2000: Current Perspective and Future Directions" (J. Bustoz et al., eds.), NATO Science Series II. Mathematics, Physics and Chemistry Vol. 30, Kluwer, Dordrecht, 2001, pp. 23–59.

[48] P. Zinn-Justin, Universality of correlation functions of Hermitian random matrices in an external field, Comm. Math. Phys. 194 (1998), 631–650.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Quasiregular Mappings, Curvature & Dynamics

Gaven J. Martin*

Abstract

We survey recent developments in the area of geometric function theory and nonlinear analysis and in particular those that pertain to recent developments linking these areas to dynamics and rigidity theory in dimension $n \geq 3$. A self mapping (endomorphism) of an n-manifold is *rational* or *uniformly quasiregular* if it preserves some bounded measurable conformal structure. Because of Rickman's version of Montel's theorem there is a close analogy between the dynamics of rational endomorphisms of closed manifolds and the classical Fatou-Julia theory of iteration of rational mappings of $\hat{\mathbb{C}}$. The theory is particularly interesting on the Riemann n-sphere $\overline{\mathbb{R}}^n$ where many classical results find their analogue, some of which we discuss here. We present the most recent results toward a solution of the Lichnerowicz problem of classifying those manifolds admitting rational endomorphisms. As a by product we discover interesting new rigidity theorems for open self maps of closed n-manifolds whose fundamental group is word hyperbolic.

Mathematics Subject Classification (2010). 30C65, 37F10, 37F30 and 30D05.

Keywords. Quasiconformal, Rational mapping, conformal dynamics.

1. Introduction

Basically a quasiregular mapping is a (possibly) branched covering map with bounded distortion. These include, for instance, piecewise linear maps between "fat" triangulations of manifolds and maps preserving measurable conformal structures as described below. The theory of quasiregular mappings - founded by Reshetnyak and Martio-Rickman-Väisälä in the 1970s - seeks to establish the analogue in higher dimensions of the geometric aspects of the theory of analytic and conformal mappings defined on subdomains of the plane, see [35, 36, 15] and

*G.J. Martin, Institute for Advanced Study, Massey University, Auckland, NZ.
E-mail: g.j.martin@massey.ac.nz

the references therein. In this regard it has been quite successful with profound applications across a spectrum of mathematics. We begin with a definition.

Definition. Let $\Omega \subset \mathbb{R}^n$ be a domain and $f : \Omega \to \mathbb{R}^n$ a mapping lying in the Sobolev space $W^{1,n}_{loc}(\Omega, \mathbb{R}^n)$ consisting of those functions which together with their first derivatives are locally n-integrable. We say f is *quasiregular* if there is a finite number K for which f satisfies the distortion inequality

$$|Df(x)|^n \leq K\, J(x, f) \tag{1}$$

where $Df(x)$ is the Jacobian matrix of f at x and $J(x, f)$ is the Jacobian determinant. Hadamard's inequality states $|Df(x)|^n \geq J(x, f)$, so the distortion inequality provides uniform bounds which are often interpreted as control of the distortion of (infinitesimal) round objects. The number K is called the *distortion* of the map f. The key nontrivial topological properties of quasiregular mappings were discovered by Reshetnyak [35],

Theorem 1.1. *A non-constant quasiregular mapping $f : \Omega \to \mathbb{R}^n$ is continuous, open and discrete.*

Here *discrete* means that the preimage of a point y, $\{x : f(x) = y\} \subset \Omega$ is discrete in Ω, that is can only accumulate on the boundary of Ω. Of course we also mean the function can be redefined on a set of measure 0 so as to have these properties.

The extension of this definition to the case of manifolds is clear. It is then a tautology that there is a measurable function $G : \Omega \to S(n)$, the non-positively curved symmetric space of positive definite symmetric $n \times n$ matrices of determinant equal to 1, for which f satisfies the Beltrami system

$$D^t f(x) Df(x) = J(x, f)^{2/n}\, G(x) \qquad \text{almost everywhere in } \Omega \tag{2}$$

By setting $G(x) = I_n$, the $n \times n$ identity matrix, whenever the left hand side of (2) is undefined, we may assume that G is defined everywhere (though the equation still only holds almost everywhere). Now such a matrix G can be used to define an inner-product on the tangent spaces to Ω by the rule

$$\langle u, v\rangle_G = \langle u, G(x)v\rangle, \qquad u, v \in T\Omega_x \tag{3}$$

In this way we view G as a *measurable conformal (or Riemannian) structure* on Ω. If $u, v \in T\Omega_x$, then almost everywhere

$$\begin{aligned} \langle f_* u, f_* v\rangle &= \langle Df(x)u, Df(x)v\rangle = \langle u, D^t f(x) Df(x) v\rangle \\ &= \langle u, J(x, f)^{2/n}\, G(x)v\rangle = J(x, f)^{2/n} \langle u, v\rangle_G \end{aligned}$$

This shows that f preserves the inner-product between tangent vectors up to a scalar multiple, therefore preserves angles, and so f can be viewed as a conformal mapping between the spaces (Ω, G) and (Ω', I_n), with $\Omega' = f(\Omega)$. As

we shall see, this is a profitable point of view. In particular (and more generally), given measurable conformal structures on domains, say (Ω, G) and (Ω', H) we can consider the families of solutions to the *generalised Beltrami systems*

$$D^t f(x)\, H(f(x))\, Df(x) = J(x,f)^{2/n}\, G(x) \qquad \text{almost everywhere in } \Omega \quad (4)$$

Typically one assumes an *ellipticity condition* on equation (4). For instance

$$\|d(G, I_n)\|_\infty + \|d(H, I_n)\|_\infty < \infty,$$

where d is the metric of $S(n)$. Basically this bounds from above and below the ratio of the largest to the smallest eigenvalues of G and H. Following the calculation above, we see that (4) implies that f is conformal between these bounded measurable structures. A few words are necessary concerning these highly nonlinear systems of partial differential equations.

2. Theory of Beltrami Systems

2.1. Two dimensions. Here it is fair to say the theory is complete and about as good as one could wish for. A thorough modern account is given in [2]. We have existence and uniqueness (up to the obvious conformal mappings) of solutions to (4) basically due to Morrey, [32]. Thanks to Astala's proof of the area distortion theorem [1], we also have optimal regularity. There is also substantial progress being made in the degenerate elliptic case [17, 2].

2.2. Rigidity $n \geq 3$. In higher dimensions the following rigidity theorem, pertaining to the case $G(x) = H(x) = I_n$, has quite important ramifications. Roughly it states that there are no conformal mappings in n-space, $n \geq 3$, apart from the obvious ones. It was first proven by Liouville in 1850 for smooth mappings [22], then refined by Gehring and Reshetnyak in the 60s and 70s, see [9, 35] to the natural space $W^{1,n}_{loc}$. However, using the nonlinear Hodge Theory developed in [16] the following sharp version is now known - sort of like the classical Looman-Menchoff Theorem in two dimensions.

Theorem 2.1. *Let $\Omega \subset \overline{\mathbb{R}}^n$, n even, be a domain and $f : \Omega \to f(\Omega)$ a $W^{1,n/2}_{loc}(\Omega, f(\Omega))$ solution to the Liouville equation*

$$D^t f(x)\, Df(x) = J(x,f)^{2/n}\, I_n \qquad \textit{almost everywhere in } \Omega \qquad (5)$$

Then there is a Möbius transformation $\Phi : \overline{\mathbb{R}}^n \to \overline{\mathbb{R}}^n$ such that

$$f = \Phi\big|\Omega$$

This result is sharp in the sense that for each domain Ω and $p < n/2$, there is a $W^{1,p}_{loc}(\Omega, f(\Omega))$ solution which is not the restriction of a Möbius transformation.

These solutions guaranteed here for $p < n/2$ are highly irregular - not even locally bounded. Conjecturally one might have this rigidity for continuous $W^{1,1}_{loc}$ solutions. For n odd, similar but less precise results are known.

2.3. Beltrami systems: Existence. Here there is very little to say apart from an obvious reworking of the classical results from the 1920s of Weyl and Schouten (which assume the vanishing of a second order tensor [46] [38]) in the case that the conformal tensors G and H are smooth. It is an extremely interesting problem to try and give reasonable conditions on G and H (even if one of them is assumed equal to I_n, the $n \times n$ Identity) which guarantee local existence if G and H are not $C^{1+\epsilon}$ smooth - most importantly the case G and H are only assumed measurable.

2.4. Beltrami systems: Uniqueness. When either $G(x)$ or $H(x)$ is equal to I_n, then an easy local application of the rigidity theorem, Theorem 2.1, together with analytic continuation (for Möbius transformations) implies global uniqueness up to a Möbius transformation. There seems no reasonable way to pose a point Cauchy problems unless G and H are highly regular. Away from these cases, there is only the following theorem known. It uses the solution to the Hilbert-Smith conjecture for quasiconformal actions on domains in space [24] and is discussed in [25].

Theorem 2.2. *Let $f, g : \Omega \to \Omega'$ be quasiconformal solutions to (4). If $f \equiv g$ on a set X of topological dimension at least $n-1$, then $f \equiv g$ on Ω. This result is sharp in the sense that the dimension of X cannot in general be lowered to $n-2$ and still have uniqueness.*

2.5. Beltrami systems: Regularity. Fairly sharp regularity results in even dimensions are known when given in terms of the operator norm of the higher dimensional Beurling transform (whose norms are not known, even in two dimensions, but can be estimated). The following two theorems of [15] give an idea of the sorts of results one can expect. The first is a slight improvement on Gehring's famous higher integrability paper [8].

Theorem 2.3. *Let $\Omega \subset \overline{\mathbb{R}}^n$ be a domain and $f : \Omega \to \overline{\mathbb{R}}^n$ a mapping. Then there is $\epsilon = \epsilon(n, G, H) > 0$ such that every $W^{1,n-\epsilon}_{loc}(\Omega, f(\Omega))$ solution to the equation (4) lies in the better space $W^{1,n+\epsilon}_{loc}(\Omega, f(\Omega))$*

The improved regularity here is important for such things as the change of variable formula and so forth. Of course ϵ here really only depends on the ellipticity constants of the equation and not directly on G and H themselves.

Next we have a topological rigidity theorem. It is basically this result which assures us we are going to have to deal with discontinuous conformal structures if there is to be a viable theory of branched mappings.

Theorem 2.4. *Let $f : \Omega \to f(\Omega)$ be a $W^{1,n}_{loc}(\Omega, f(\Omega))$, $n \geq 3$, solution to the equation (4) where both G and H are continuous. Then f is a local homeomorphism.*

In fact rather more can be said here. The result holds if G and H are only close to continuous in a BMO sense. Thus to admit branching, we must have jump discontinuities in G or H of a fixed size (in the metric of $S(n)$).

3. Uniformly Quasiregular Mappings

Having now much of the basic theory at hand we discuss connections with dynamics and in particular analogues of the theory of iteration of rational mappings of $\hat{\mathbb{C}}$.

A quasiregular map $f : \overline{\mathbb{R}}^n \to \overline{\mathbb{R}}^n$ with a uniform bound on the distortion of all its iterates $f \circ f, \ldots, f \circ \cdots \circ f, \ldots$ is called *uniformly quasiregular* (uqr). Such maps are always rational with respect to some measurable Riemannian structure [15]. This means that there is a bounded measurable $G : \overline{\mathbb{R}}^n \to S(n)$ such that

$$D^t f(x) G(f(x)) Df(x) = J(x, f)^{2/n} G(x), \qquad a.e. \ \ \overline{\mathbb{R}}^n \tag{6}$$

The space of $W^{1,n}(\overline{\mathbb{R}}^n)$ solutions to this nonlinear PDE forms a semigroup analogous to the analytic functions – and quasiconformally conjugate to the rational functions in two-dimensions. Because of Rickman's version of Montel's Theorem [36] there is a reasonably complete Fatou-Julia theory associated with the iteration of uqr mappings, this was started in joint work with Iwaniec [15], but has been developed by V. Mayer, K. Peltonen and others, see [26, 27]. There are also strong restrictions on the geometry and topology of closed manifolds admitting nontrivial uqr mappings, for instance (as we shall see) they cannot be negatively curved.

The Fatou set $\mathcal{F}(f)$ of a uqr-mapping f is the open set where the iterates form a normal family (that is have locally uniformly convergent subsequences). The Julia set $\mathcal{J}(f)$ is the complement of the Fatou set

$$\mathcal{J} = \overline{\mathbb{R}}^n \setminus \mathcal{F}$$

If the degree of $f \geq 2$, the only interesting case for us, then the Julia set is nonempty, closed and a completely invariant set,

$$f^{-1}(\mathcal{J}) = \mathcal{J}$$

The following factorisation theorem shows that in fact uqr mappings are quite common.

4. Stoilow Factorisation

We have the following variant of Stoïlow's theorem, [28].

Theorem 4.1. *Suppose $g : \overline{\mathbb{R}}^n \to \overline{\mathbb{R}}^n$ is a non-constant quasiregular mapping, $n \geq 2$. Then there exists a uniformly quasiregular mapping f whose Julia set is a Cantor set, and a quasiconformal mapping $h : \overline{\mathbb{R}}^n \to \overline{\mathbb{R}}^n$ such that $g = f \circ h$.*

Indeed it is shown that the uqr mapping is structurally stable (or generic), there is a single attracting fixed point, no relations between critical points and the Julia set is ambiently quasiconformally equivalent to the middle thirds Cantor set (so is not wild).

Classically the factorization (for quasiregular maps of $\hat{\mathbb{C}} = \mathbb{S}^2$) is unique up to Möbius transformation. If $\varphi \circ f = \psi \circ g$, then there is a Möbius transformation Φ so that $\varphi \circ \Phi = \psi$. Clearly this statement cannot hold in higher dimensions if φ and ψ are merely assumed uqr. However if we fix the invariant conformal structure, then we can make uniqueness statements up to a finite dimensional Lie group.

Theorem 4.2. *Let G be bounded measurable conformal structure on $\overline{\mathbb{R}}^n$. Then there is a closed finite dimensional Lie group Γ of quasiconformal homeomorphisms of $\overline{\mathbb{R}}^n$ with the following property: If $g, h : \overline{\mathbb{R}}^n \to \overline{\mathbb{R}}^n$ are quasiconformal mappings such that*

$$\varphi \circ g = \psi \circ h : \overline{\mathbb{R}}^n \to \overline{\mathbb{R}}^n, \tag{7}$$

and both φ and ψ are rational with respect to G, then there is $\gamma \in \Gamma$ such that

$$\varphi \circ \gamma = \psi : \overline{\mathbb{R}}^n \to \overline{\mathbb{R}}^n$$

We remark that in two dimensions the space of generic uqr mappings that our factorization produces can be described.

4.1. Smooth uqr mappings. The technique used in construction of the factorisation is sufficiently robust that if the map f is smooth of class $C^k(\overline{\mathbb{R}}^n)$, then the quasiconformal homeomorphism h, and consequently the uqr mapping φ, can be chosen to be smooth of the same class. Typically one does not expect branched (not locally injective) quasiregular mappings to be smooth, however there are examples of M. Bonk and J. Heinonen [3] of a quasiregular map $f : \mathbb{S}^3 \to \mathbb{S}^3$ which is $C^{3-\epsilon}(\mathbb{S}^3)$ for every $\epsilon > 0$. Kaufman, Tyson and Wu extended these results to higher dimensions, [19]. The following theorem (which was certainly known to Bonk and Heinonen) is a consequence.

Theorem 4.3. *There are smooth uqr mappings of $\overline{\mathbb{R}}^n$ with nonempty branch set, $B_f \neq \emptyset$. Indeed,*

- *For each $\epsilon > 0$, there is a $C^{3-\epsilon}(\mathbb{S}^3)$ uniformly quasiregular mapping φ whose Julia set is a Cantor set.*

- *For each $\epsilon > 0$, there is a $C^{2-\epsilon}(\mathbb{S}^4)$ uniformly quasiregular mapping φ whose Julia set is a Cantor set.*
- *For each $n \geq 5$ there is an $\epsilon = \epsilon(n) > 0$ and a $C^{1+\epsilon}(\mathbb{S}^n)$ uniformly quasiregular mapping φ whose Julia set is a Cantor set.*

Note that although these maps are smooth, any invariant conformal structure must be quite irregular near the branch set and the Julia set.

5. Dynamics of UQR mappings

We first consider the classification of fixed points of uqr-mappings. In [14] we showed that uniformly quasiregular mappings are locally Lipschitz near a fixed point x_0 which is not a branch point. This is used to show that the family $\mathcal{F} = \{f_\lambda \colon \lambda > 1\}$ is a normal family, where $f_\lambda(z) = \lambda f(z/\lambda)$. Moreover, it is shown that all limits of convergent subsequences of $\mathcal{F}$ are uniformly quasiconformal mappings. The set of all such limit mappings is called the generalized derivative of f at x_0. Uniformly quasiconformal mappings have been classified as either loxodromic, elliptic or parabolic. It follows that the elements of the generalized derivative are either all constant, all elliptic, or all loxodromic, and this allows for a classification of the fixed points of a uniformly quasiregular mapping as attracting/repelling (generalized derivative loxodromic), neutral (generalized derivative elliptic), or super-attracting (fixed point is a branch point).

Theorem 5.1. *The Fatou set of a uniformly quasiregular mapping has precisely the same types of stable components as rational functions do.*

For attracting and repelling fixed points we know

Theorem 5.2. *Any uniformly quasiregular mapping is locally quasiconformally conjugate to the Möbius group generated by $x \mapsto 2x$ near a repelling fixed point. Any uniformly quasiregular mapping is quasiconformally conjugate to $x \mapsto \frac{1}{2}x$ near an attracting fixed point.*

There are examples of uqr maps with parabolic dynamics. It can be shown that a map with a parabolic fixed point can be constructed in such a way that it does not admit a quasiconformal linearization in its attracting parabolic petal (unlike the rational case).

Question: A very interesting problem is to decide whether or not it is possible to have a "Siegel disk" (presumably a ball or solid torus with irrational rotational dynamics) for a non-injective uqr mapping.

A classical result is the density of repellors in the Julia set. This is not known in complete generality yet for uqr mappings. We do know

Theorem 5.3. *The set of repelling and neutral fixed points is dense in the Julia set.*

In certain cases (assumptions on the topological structure of the Julia set such as separating) we do know repellors are dense.

6. Rational Maps of Manifolds

A natural question is to ask what sort of manifolds support rational endomorphisms. In two dimensions it is an easy application of the signature formula for branched coverings to see that only the sphere and torus admit noninjective rational self maps. In higher dimensions the question is bound to be more complicated - though a complete answer was given by Kangaslampi in three dimensions [18].

6.1. The Lichnerowicz problem. Here we shall consider such mappings acting on *closed* manifolds M of dimension at least two and our problem is to determine what kind of manifolds admit the action of such a map and also to determine what kind of uqr mappings can act on a given manifold. A result of Sullivan (see Tukia - Väisälä [42]) shows that any topological n-manifold, $n \neq 4$, admits a unique quasiconformal structure and this is enough to define a bounded measurable Riemannian structure and what it means to be quasiregular [36]. In four dimensions (where there are many interesting and unsolved questions from the point of view of quasiconformal geometry) we shall have to suppose our manifolds admit such a structure.

Note that a uqr map $f : M \to M$ is surjective since the continuity and openness of a quasiregular map implies that the image fM is both compact and open; hence $fM = M$. The first part of our problem is a non-injective version of the answer given by Ferrand [7] to a conjecture of Lichnerowicz [21]. She essentially showed that, up to quasiconformal equivalence, the only compact n-manifold which admits a non-compact quasiconformal group action is the standard n-sphere $\mathbb{S}^n$. If there is a uqr map f of a closed manifold M, then the semi-group $\{f^n\}_{n=1}^{\infty}$ is non-compact (in fact the Julia set of f is always non-empty). Therefore, the existence of such a map should imply severe restrictions on the manifold M. The first of these is the following obstruction for the existence of uqr maps, [27].

Theorem 6.1. *If M^n is a closed n-manifold and $f : M^n \to M^n$ is a non-injective uqr mapping, then there exists a non-constant quasiregular mapping $g : \mathbb{R}^n \to M^n$.*

Manifolds admitting such a map g are called *qr-elliptic*, and answering a question of Gromov, Varopoulos, Saloff-Coste and Coulhon [44] showed that M must in turn have a fundamental group of at most polynomial growth.

The generalized Lichnerowicz problem seeks to determine determining all closed manifolds which admit non-injective uqr mappings: This problem was discussed and largely solved in [27] with the following results.

Theorem 6.2. *Let f be a non-injective uqr map of the closed manifold M and suppose that f is locally homeomorphic, that is, the branch set $B_f = \emptyset$. Then M is the quasiconformal image of a Euclidean space form.*

and as a sort of converse

Theorem 6.3. *. If M is quasiconformally equivalent to a euclidean space form, then M admits no branched quasiregular (and in particular no branched uqr) mappings.*

Actually under these circumstances we prove that M does not admit an orientation preserving, discrete and open map which is branched. In the case of the sphere, lens spaces and other spherical manifolds the existence of uqr maps is due to [16] and K. Peltonen [34]. These results suggest that there are few uqr mappings in three or more dimensions compared with the space of rational functions of the Riemann sphere $\mathbb{S}^2 = \hat{\mathbb{C}}$.

By a Euclidean space form we mean the quotient of $\mathbb{R}^n$ under a Bieberbach group (co-finite volume lattice) $\Gamma \subset Isom^+(\mathbb{R}^n)$. The two other types of space forms are the quotients by torsion free co-finite volume lattices of isometries of the n-sphere and of hyperbolic n-space. In these cases we have

For Euclidean space forms we gave a complete description of the possible uqr mappings:

Theorem 6.4. *Any non-injective uqr map of a closed Euclidean space form M is the quasiconformal conjugate of a conformal map.*

We remark that, in this second result, we no longer suppose that the map has to be locally injective. This result is surprising because it is false for globally injective mappings. Indeed, V. Mayer shows that there are uniformly quasiconformal (even bi-Lipschitz) maps of three (or higher) dimensional tori which cannot be quasiconformally conjugate to a conformal map [23]. Next, we distinguish space forms according to the type of uqr maps they support:

Theorem 6.5. *If M is a closed space form, then we have the following characterization:*

1. *M admits a branched uqr map if and only if M is a spherical space form.*
2. *M admits a non-injective, locally injective uqr map if and only if M is a Euclidean space form.*
3. *M admits no non-injective uqr map if and only if M is a hyperbolic space form.*

We suggest that the correct counterpart for the Lichnerowicz conjecture for uqr maps could be that if a closed manifold M supports a non-injective uqr map, then it must be a finite product

$$M = M_1 \times \cdots \times M_k$$

of closed manifolds M_i which are quasiconformal images of either Euclidean or spherical space forms. It remains an interesting open question whether for example a space like $(\mathbb{S}^2 \times \mathbb{S}^2)\#(\mathbb{S}^2 \times \mathbb{S}^2)$ supports uqr maps. This manifold was very recently proven to be elliptic by S. Rickman [37] and perhaps represents a best candidate as a counterexample to the conjecture.

6.2. Negative curvature. Here we study the existence or otherwise of branched (not locally injective) quasiregular maps between manifolds of negative curvature as discussed in work with M. Bridson and A. Hinkannen [4]. Our main results are to show, as a particular case of a more general result, that a branched quasiregular mapping $f : M \to N$ between closed hyperbolic manifolds can never induce an injection on fundamental groups. This also strengthens the earlier results of [27] discussed above - that the only *uniformly* quasiregular automorphisms of closed hyperbolic manifolds are the obvious ones (i.e., uniformly quasiconformal mappings isotopic to periodic isometries). In fact in [4] we prove that such manifolds admit no non-obvious quasiregular self mappings at all - there are no discrete open self maps whatsoever which are not homeomorphisms isotopic to an isometry. We give a number of related results, including an extension of the above theorem to convex co-compact manifolds and a generalization concerning open mappings between closed negatively curved manifolds of dimension $n \neq 4$. We further discuss the case of complete finite-volume hyperbolic manifolds. The proofs of these latter results rely on an analysis of the self-maps of word hyperbolic and relatively hyperbolic groups which are of independent interest.

6.3. Proper open surjections and π_1. The following lemma of Walsh [45] and Smale [41] will turn out to be quite important in what follows. Recall a map is *proper* if the preimage of a compact set is compact.

Lemma 6.6. *Let M_1 and M_2 be connected manifolds (possibly with boundary). If a map $f : M_1 \to M_2$ is proper, open and surjective, then the index of $f_*\pi_1 M_1$ in $\pi_1 M_2$ is finite.*

The same argument applies with minor modification in the case that M_i are orbifolds. Further it is also noted in [41] Theorem 3 that under the hypotheses of the lemma the map f induces a surjection on rational homology. Finally, in light of what is to follow we note a main result of Walsh's paper, Corollary 5.15.3, essentially a converse to Lemma 6.6

Theorem 6.7. *If M and N are compact connected PL manifolds and $f : M \to N$ a map with $|f_*\pi_1 M_1 : \pi_1 M_2| < \infty$, then f is homotopic to a light open mapping*

Here a *light* mapping is one for which the preimage of every point is totally disconnected, for instance a Cantor set.

7. Quasiregular Mappings Between Hyperbolic Manifolds

Chernavski [5] and Väisälä [43] proved a key theorem stating that a discrete open mapping $f : M \to M$ of an n-manifold has the dimension of its branch set less than or equal to $n-2$, $\dim(B_f) \leq n-2$, and further $\dim(f(B_f)) \leq n-2$. As a consequence of the Hurewitz and Wallman theorems, the set B_f does not locally separate M at any point. Thus, $f|M \setminus f^{-1}(f(B_f)) \to M \setminus f(B_f)$ is a covering map, where $M \setminus f^{-1}(f(B_f))$ itself is a connected open manifold, dense in M. Moreover, for each $y \in M \setminus f(B_f)$ the set $f^{-1}(y)$ must contain exactly the same number d (the *degree*) points.

If $\Gamma \subset \mathrm{Isom}^+(\mathbb{H}^n)$ is a discrete non-elementary (*Kleinian*) group, we denote its limit set by $\Lambda(\Gamma) \subset \partial\mathbb{H}^n = \overline{\mathbb{R}}^{n-1}$. The orbit space of a Kleinian group Γ is $\mathbb{H}^n/\Gamma$ a hyperbolic orbifold (or manifold should the group Γ be torsion free). In what follows dim refers to the topological dimension while $\dim_H$ refers to the Hausdorff dimension.

Theorem 7.1. *For $i = 1, 2$, let M_i be a hyperbolic n–orbifold with fundamental group Γ_i and limit sets Λ_i. Let $f : M_1 \to M_2$ be a proper quasiregular mapping such that $f_* : \Gamma_1 \to \Gamma_2$ has finite kernel. Suppose that one of the following conditions is satisfied:*

1. $\dim(\Lambda_1) \geq n-2$
2. $\dim(\Lambda_2) \geq n-2$
3. $\dim_H(\Lambda_1) = n-1$
4. $\dim_H(\Lambda_2) = n-1$

Then f is a finite-sheeted covering map whose lift to $\mathbb{H}^n$ is a quasiconformal homeomorphism with $f(\Lambda_1) = \Lambda_2$.

If the hyperbolic manifolds M_i are closed, then the hypothesis that f is proper is redundant and all four conditions are satisfied. Thus

Corollary 7.2. *Let M and N be closed hyperbolic n-manifolds. Then there is no π_1-injective branched quasiregular mapping $f : M \to N$.*

Notice next that the above dimension hypothesis on the limit sets is satisfied if one of them separates. This will be the case if, for instance, one of the manifolds has more than one boundary component (not a cusp). For instance if, say M is convex co-compact - i.e. $(\overline{\mathbb{H}}^n \setminus \Lambda_M)/\Gamma_M$ is compact where the boundary components are the manifolds $(\partial\mathbb{H}^n \setminus \Lambda_M)/\Gamma_M$.

Corollary 7.3. *Let M and N be convex co-compact hyperbolic n-manifolds one of which has more than one boundary component. Then there is no proper π_1-injective branched quasiregular mapping $f : M \to N$.*

It is possibly true that more generally if M and N are hyperbolic manifolds, then there is no proper branched π_1-injective quasiregular mapping. In three dimensions this is true.

Theorem 7.4. *Let M and N be hyperbolic 3-manifolds. Then there is no proper π_1-injective branched quasiregular mapping $f : M \to N$.*

Application of the conditions on the Hausdorff dimension of the limit sets can be found when considering geometrically infinite Kleinian groups.

It is clear from Theorem 7.1 that we will have to address the question of when a quasiregular map, or more generally an open map, between hyperbolic manifolds induces an injection on fundamental groups.

8. Endomorphisms of Hyperbolic Groups

The following theorem was proved by Z. Sela, [39], Theorem 3.9 (also [40], Theorem 1.12) in the course of his work on the Hopf property for word hyperbolic groups.

Theorem 8.1. *Let Γ be a torsion-free hyperbolic group and let $\phi : \Gamma \to \Gamma$ be a homomorphism. Then there is an integer k_0, so that $\ker(\phi^n) = \ker(\phi^{k_0})$ for every $n > k_0$.*

Sela also proved that torsion-free, freely indecomposable, non-elementary word hyperbolic groups are co-Hopfian, [40]. In particular co-compact lattices of n-dimensional hyperbolic space are co-Hopfian. Daniel Groves recently extended Sela's results from the hyperbolic setting to toral relatively hyperbolic groups [13]. Every geometrically-finite subgroup of $\mathrm{SO}(n,1)$ has a subgroup of finite index that lies in this class.

In the current setting we require a variation on this: we must allow freely-decomposable groups, but we need only constrain homomorphisms whose image is of finite index.

Lemma 8.2. *If a finitely generated torsion-free group Γ can be expressed as a free product $\Gamma = A * B$ with A and B nontrivial, then there does not exist an injective homomorphism $\phi : \Gamma \to \Gamma$ with $1 < [\Gamma : \phi(\Gamma)] < \infty$.*

These results now combine to give us the following interesting result of [4].

Theorem 8.3. *If Γ is a torsion-free non-elementary hyperbolic group, then there is no homomorphism $\phi : \Gamma \to \Gamma$ with $1 < [\Gamma : \phi(\Gamma)] < \infty$.*

For residually finite groups (such as subgroups of $\mathrm{SO}(n,1)$, our main interest) one also deduce Theorem 8.3 from Theorem 8.1 by using the following generalisation of Malcev's famous observation that finitely generated residually finite groups are Hopfian.

Proposition 8.4. *If Γ is finitely-generated, torsion-free and residually finite, and $\phi : \Gamma \to \Gamma$ is a homomorphism for which $[\Gamma : \phi^k(\Gamma)]$ remains bounded as $k \to \infty$, then ϕ is an isomorphism.*

8.1. Topological rigidity results.

The following is an immediate consequence of Theorem 8.3 and Lemma 6.6.

Theorem 8.5. *If M is a connected manifold whose fundamental group is torsion-free, non-elementary and word hyperbolic, then every proper open surjective mapping $f : M \to M$ induces an isomorphism of the fundamental group.*

Remark 8.6. We stated Theorem 8.5 only for manifolds for simplicity. But the proof of Walsh's Lemma applies in far greater generality and hence the above theorem can be generalised enormously: it suffices to assume that M is a locally-finite cell complex, for example. Further, if M is closed then the hypotheses that f is surjective and proper are redundant.

Corollary 8.7. *If M is a closed n-manifold whose fundamental group is torsion-free, non-elementary and word hyperbolic, then every quasiregular mapping $f : M \to M$ is a homeomorphism.*

Proof. Every quasiregular map $f : M \to M$ is open discrete and so finite to one. As above, the induced map on fundamental group is an isomorphism. Thus has f has degree 1, $B(f) = \emptyset$ and f is a covering map by [43]. Thus f is a homeomorphism. □

The Farrell and Jones [6] topological rigidity theorem for non-positively curved manifolds tells us that closed negatively curved manifolds of dimension $n \geq 5$ are homeomorphic if their fundamental groups are isomorphic. Perelman's proof of the geometrisation conjecture implies that the same result is true in dimension 3. If the curvature is strictly negative, the fundamental group of such a manifold is word hyperbolic. Thus Theorem 8.5 implies:

Theorem 8.8. *Let M_1 and M_2 be closed Riemannian n-manifolds ($n \neq 4$) of negative sectional curvature. Suppose there are open maps $f : M_1 \to M_2$ and $g : M_2 \to M_1$. Then M_1 is homeomorphic to M_2.*

Proof. By Theorem 8.5, the compositions $f \circ g : M_2 \to M_2$ and $g \circ f : M_1 \to M_1$ induce isomorphisms on π_1. □

Sela noted a version of this result for degree 1 maps. Ian Agol has suggested an alternative proof in the hyperbolic case based on the Gromov norm.

The results of Farrell-Jones and Perelman also yield the refinement:

Theorem 8.9. *If M is a closed n-manifold of negative sectional curvature ($n \geq 5$), then every open mapping $f : M \to M$ is homotopic to a homeomorphism.*

9. Quasiregular Maps and Rigidity of Hyperbolic Manifolds

Putting together these results gives the following.

Theorem 9.1. *If M is a convex co-compact hyperbolic n–manifold, then every proper, quasiregular mapping $f : M \to M$ is a homeomorphism.*

Proof. Once again Lemma 6.6 tells us that $f_*(\pi_1 M) \subset \pi_1 M$ is of finite index, and Theorem 8.3 then tells is that f_* is an isomorphism. Theorem 7.1 then tells us that f is a homeomorphism. □

We combine what the above result with Mostow rigidity to obtain the first item in the following corollary, and include the Euclidean case for comparison.

Corollary 9.2. *Let M be a space form and $f : M \to M$ quasiregular.*

- *If M is hyperbolic, then f is quasiconformal and homotopic to an isometry of finite period.*
- *If M is euclidean, then f is quasiconformally conjugate to a multiplication or f is quasiconformal (i.e., injective).*

In the spherical case, above dimension 1, a locally injective map must be injective.

Our results combine to give the following generalisation of the Mostow rigidity theorem [33],

Theorem 9.3. *Let M_1 and M_2 be closed hyperbolic n–manifolds. Suppose there is an open mapping $f : M_1 \to M_2$ and an injection $\phi : \pi_1(M_2) \to \pi_1(M_1)$ with $[\pi_1(M_1) : \phi(\pi_1(M_2))] < \infty$. Then f is homotopic to an isometry.*

One can extend our results for convex co-compact subgroups of $SO(n, 1)$ to all geometrically finite torsion-free lattices by using the work of Groves [13] in place of Sela's Theorem. Walsh's Lemma still applies in this context, but one needs an adaptation of Theorem 7.1, which we do not present here. The conclusion is that every quasiregular self mapping of a finite volume hyperbolic n–manifold is isotopic to an isometry.

Finally we want to make the following observation relating what we have above with Wilson's counterexample to the Whyburn conjecture [47] giving a dichotomy between discrete open and light open mappings of hyperbolic manifolds.

Theorem 9.4. *Let M be a closed hyperbolic n–manifold. Then any discrete open mapping $f : M \to M$ is a homeomorphism isotopic to an isometry. However, there are light open mappings $g : M \to M$ which are not homeomorphisms.*

Presumably the arguments used to prove these results lead to the same conclusion if M is a closed negatively curved n–manifold whose universal cover has boundary an $(n-1)$-sphere. The action of the fundamental group (Gromov hyperbolic) on the boundary sphere as a convergence group [10] with every limit point a point of approximation (or conical limit point) will imply that the induced boundary map of the lift of f is a homeomorphism (see [29, 30] for proofs of these sorts of results) since we know the induced map on fundamental groups is an isomorphism.

References

[1] K. Astala, *Area distortion of quasiconformal mappings*, Acta Math., **173**, (1994), 37–60.

[2] K. Astala, T. Iwaniec and G.J. Martin, *Elliptic partial differential equations and quasiconformal mappings in the plane*, Princeton Mathematical Series, 48. Princeton University Press, Princeton, NJ, 2009.

[3] M. Bonk, and J. Heinonen, *Smooth quasiregular mappings with branching*, Publ. Math. Inst. Hautes tudes Sci., **100**, (2004), 153–170.

[4] M. Bridson, A. Hinkkanen and G. Martin, *Quasiregular self-mappings of manifolds and word hyperbolic groups*, Compos. Math., **143**, (2007), 1613–1622.

[5] A.V. Cernavskiĭ, *Finite-to-one open mappings of manifolds.*, (Russian) Mat. Sb. (N.S.), **65**, (107), 1964, 357–369.

[6] T. Farrell and L. Jones, *Compact negatively curved manifolds (of dim$\neq 3,4$) are topologically rigid*, Proc. (US) Nat. Acad. Sci., **86**, (1989), 3461-3463.

[7] J. Ferrand, *Transformations conformes et quasi-conformes des vari?et?es riemanniennes compactes (démonstration de la conjecture de A. Lichnerowicz)*, Acad. Roy. Belg. Cl. Sci. Mém. Coll., **39**, 44 pp. (1971).

[8] F.W. Gehring, *The L^p-integrability of the partial derivatives of a quasiconformal mapping*, Acta Math., **130**, (1973), 265–277

[9] F.W. Gehring, *Rings and quasiconformal mappings in space*, Trans. Amer. Math. Soc., **103**, (1962), 353–393.

[10] F.W. Gehring and G.J. Martin, *Discrete quasiconformal groups, I*, Proc. London Math. Soc., (3), **55**, (1987), 331–358.

[11] M. Gromov, *Structures Métriques pour les Variétés Riemanniennes*, J. Lafontaine and P. Pansu Eds. Cedric/Fernand, Paris, 1981.

[12] Gromov M., Hyperbolic groups. In *Essays in Group Theory* (S.M. Gersten, ed.), Springer-Verlag, New York, 1987, pp. 75–263.

[13] D. Groves, *Limit groups for relatively hyperbolic groups, II: Makanin-Razborov diagrams*, math.GR/0503045.

[14] A. Hinkkanen, G.J. Martin and V. Mayer *Dynamics of uniformly quasiregular mappings*, Math. Scand. **95**, (2004), no. 1, 80–100.

[15] T. Iwaniec and G.J. Martin, *Quasiregular Semigroups*, Ann. Acad. Sci. Fenn. Math. **21** (1996) 241–254

[16] T. Iwaniec and G.J. Martin, *Quasiregular mappings in even dimensions*, Acta Math., **170**, (1993), 29–81.

[17] T. Iwaniec and G.J. Martin, *The Beltrami Equation*, Memoirs of the Amer. Math. Soc., **191**, (2008), no. 893, x+92 pp.

[18] R. Kangaslampi, *Uniformly quasiregular mappings on elliptic Riemannian manifolds*, Ann. Acad. Sci. Fenn. Math. Diss. **151**, 2008.

[19] R. Kaufman, J. Tyson and J-M Wu, *Smooth quasiregular maps with branching in* $\mathbb{R}^n$, Publ. Math. Inst. Hautes tudes Sci., **101**, (2005), 209–241.

[20] J. Lelong–Ferrand, *Invariants conformes globaux sur les varietes riemanniennes*, J. Diff. Geom., **8**, (1973), 487–510.

[21] A. Lichnerowicz, *Sur les transformations conformes d'une variété riemannienne compacte*, C. R. Acad. Sci. Paris **259**, (1964), 697–700.

[22] J. Liouville, *Théorèm sur l'équation* $dx^2 + dy^2 + dz^2 = \lambda(d\alpha^2 + d\beta^2 + d\gamma^2)$, J. Math. Pures Appl., **1**, (15) (1850), 103.

[23] V. Mayer, *Cyclic parabolic quasiconformal groups that are not the quasicoformal conugates of Möbius groups*, Ann. Acad. Sci. Fenn. Ser. AI Math., **18**, (1993), 147–154.

[24] G.J. Martin *The Hilbert-Smith conjecture for quasiconformal actions*, Electron. Res. Announc. Amer. Math. Soc., **5** (1999), 66–70

[25] G.J. Martin *Analytic continuation for Beltrami systems, Siegel's theorem for UQR maps, and the Hilbert-Smith conjecture*, Math. Ann., **324**, (2002), 329–340.

[26] G.J. Martin and V. Mayer *Rigidity in holomorphic and quasiregular dynamics*, Trans. Amer. Math. Soc., **355**, (2003), 4349–4363.

[27] G.J. Martin, V. Mayer and K. Peltonen, *The generalized Lichnerowicz problem: uniformly quasiregular mappings and space forms*, update in Proc. Amer. Math. Soc. **134**, (2006), 2091–2097.

[28] G.J. Martin and K. Peltonen, *Stoïlow factorization for quasiregular mappings in all dimensions*, Proc. Amer. Math. Soc., **138**, (2010), 147–151.

[29] G.J. Martin and P. Tukia, *Convergence and Mbius groups*, Holomorphic functions and moduli, Vol. II (Berkeley, CA, 1986), 113–140, Math. Sci. Res. Inst. Publ., 11, Springer, New York, 1988.

[30] G.J. Martin and P. Tukia, *Convergence groups with an invariant component pair*, Amer. J. Math., **114**, (1992), 1049–1077.

[31] V. Mayer, *Uniformly quasiregular mappings of Lattés type*, Conform. Geom. Dyn., **1**, (1997), 104–111.

[32] C. B. Morrey, *On the solutions of quasi-linear elliptic partial differential equations*, Trans. Amer. Math. Soc., **43**, (1938), 126–166.

[33] G.D. Mostow, *Quasi-conformal mappings in n-space and the rigidity of hyperbolic space forms,* Inst. Hautes Études Sci. Publ. Math., **34**, (1968), 53–104.

[34] K. Peltonen, *Examples of uniformly quasiregular mappings*, Conformal Geom. Dyn. **3**, (1999), 158–163.

[35] Yu. G. Reshetnyak, *Space mappings with bounded distortion*, Translated from the Russian by H. H. McFaden. Translations of Mathematical Monographs, 73, American Mathematical Society, Providence, RI, 1989

[36] S. Rickman, *Quasiregular mappings*, Springer, 1993.

[37] S. Rickman, *Simply connected quasiregularly elliptic 4-manifolds*, Ann. Acad. Sci. Fenn. Math., **31**, (2006), 97–110.

[38] J.A. Schouten, *Der Ricci-Kalkl.* (German) [Ricci calculus] Eine Einfhrung in die neueren Methoden und Probleme der mehrdimensionalen Differentialgeometrie. Reprint of the 1924 original. Grundlehren der Mathematischen Wissenschaften, 10. Springer-Verlag, Berlin-New York, 1978. x+312 pp.

[39] Z. Sela, *Endomorphisms of hyperbolic groups:* **I**, Topology, **38**, (1999), 301-321.

[40] Z. Sela, *Structure and rigidity in (Gromov) hyperbolic groups and discrete groups in rank* 1 *Lie groups. II.*, Geom. Funct. Anal., **7**, (1997), 561–593.

[41] S. Smale, *A note on open maps*, Proc. Amer. Math. Soc., **8**, (1957), 391–393.

[42] P. Tukia and J. Väisälä, Lipschitz and quasiconformal approximation and extension., Ann. Acad. Sci. Fenn. Ser. A I Math., **6**, (1981), 303–342 (1982)

[43] J. Väisälä, *Discrete open mappings on manifolds*, Ann. Acad. Sci. Fenn. Ser. A I, **392**, 1966 10 pp

[44] N.Th. Varopoulos, L. Saloff-Coste and T. Coulhon, *Analysis and geometry on groups*, Cambridge Univ. Press, Cambridge, 1992.

[45] J.J. Walsh, *Light open and open mappings on manifolds,* **II**, Transactions of the Amer. Math. Soc., **217**, (1976), 271–284.

[46] H. Weyl, *Zur Infinitesimalgeometrie; p dimensionale Flche im n dimensionalen Raum.* (German) Math. Z. **12**, (1922), 154–160

[47] D. Wilson, *Open mappings on manifolds and a counterexample to the Whyburn conjecture*, Duke Math. J., **40**, (1973), 705–716.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Random Complex Zeroes and Random Nodal Lines

Fedor Nazarov* and Mikhail Sodin†

To the memory of Oded Schramm

Abstract

In these notes, we describe the recent progress in understanding the zero sets of two remarkable Gaussian random functions: the Gaussian entire function with invariant distribution of zeroes with respect to isometries of the complex plane, and Gaussian spherical harmonics on the two-dimensional sphere.

Mathematics Subject Classification (2010). 30B20, 33C55 and 60G55.

Keywords. Gaussian entire functions, random complex zeroes, random waves, random nodal lines.

These notes consist of two almost independent parts. In both of them, we talk about zeroes of special Gaussian random functions. To understand them, we had to combine various tools from complex and real analysis with rudimentary probabilistic methods. We think that the results and techniques presented here can serve as guidelines in other problems of similar nature arising in analysis, mathematical physics, and probability theory.

The function F that we consider in the first part is a random analytic function of one complex variable. In this case, one can recover the zeroes of F by applying the Laplacian to $\log|F|$. This paves the way for using complex analysis tools, and for this reason, the problems that we discuss in the first part are pretty well understood by now, though some intriguing questions still remain open.

In the second part, we deal with topological properties of the zero sets of random (real-valued) functions of several real variables. This is an area with

*F.N.: Mathematics Department, University of Wisconsin-Madison, 480 Lincoln Dr., Madison WI 53706 USA. E-mail: nazarov@math.wisc.edu.

†M.S.: School of Mathematics, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: sodin@post.tau.ac.il.

This work is partially supported by grant No. 2006136 of the United States - Israel Binational Science Foundation.

wealth of interesting and difficult questions and with very few advances. In essence, in this part, the reader will find a discussion of one recent theorem on the number of connected components of zero sets of Gaussian spherical harmonics along with various open questions.

Part I. Random Complex Zeroes

The study of zeroes of random polynomials and random analytic functions has a long history. It started with the pioneering works of Kac, Littlewood, Offord, Rice, and Wiener, and was later continued by Hammersley, Kahane, Maslova, and many others. The subject was revived in the 1990's by several groups of researchers (Bogomolny-Bohigas-Leboeuf, Shub-Smale, Edelman-Kostlan, Ibragimov-Zeitouni, Hannay, Bleher-Shiffman-Zelditch, Nonnenmacher-Voros) who came from very different areas and established new links to mathematical physics, probability theory, and complex geometry. Some of these results were surveyed in the lectures by Zelditch [54] and Sodin [44]; see also an introductory article [33] and the recent book by Hough, Krishnapur, Peres, and Virág [21].

In particular, Kostlan, Bogomolny-Bohigas-Leboeuf, Shub-Smale, and Hannay introduced a remarkable construction of random Gaussian entire functions with translation invariant distribution of their zeroes. Let

$$F(z) = \sum_{n\geq 0} \zeta_n \frac{z^n}{\sqrt{n!}}$$

where ζ_n are independent standard complex Gaussian random coefficients (i.e., the density of ζ_k with respect to the Lebesgue measure in $\mathbb{C}$ is $\frac{1}{\pi}e^{-|\zeta|^2}$). The distribution of the random function F is invariant with respect to rotations around the origin, but it is *not* translation invariant, for instance, because $\mathcal{E}|F(z)|^2 = e^{|z|^2}$ (here and below, $\mathcal{E}$ means the mathematical expectation). However, the distribution of the zero set $\mathcal{Z} = F^{-1}\{0\}$ *is* translation invariant. One of the ways to see this is to check that the Gaussian random function

$$F_\lambda(z) = F(z+\lambda)e^{-z\overline{\lambda}-\frac{1}{2}|\lambda|^2}, \qquad \lambda \in \mathbb{C},$$

has the same covariance function as F:

$$\mathcal{E}\{F(z)\overline{F(w)}\} = \mathcal{E}\{F_\lambda(z)\overline{F_\lambda(w)}\} = e^{z\overline{w}}$$

which is nothing else but the reproducing kernel in the classical Fock-Bargmann space of entire functions. This coincidence is not accidental [33]. Moreover, due to remarkable Calabi's rigidity [21, Section 2.5], this is the *only* translation invariant zero set of a Gaussian entire function up to scaling. We call the function F the Gaussian Entire Function (G.E.F., for short).

It is worth mentioning that there exist similar constructions for other domains with transitive groups of isometries (the hyperbolic plane, the Riemann sphere, the cylinder and the torus).

1. Linear Statistics

One of the most traditional ways to study asymptotic properties of a random discrete subset X of the plane is to take a test-function h, and to look at the asymptotic behaviour of the linear statistics

$$n_X(r,h) = \sum_{a\in X} h\big(\tfrac{a}{r}\big)$$

as $r\to\infty$. We put $n(r,h) = n_{\mathcal{Z}}(r,h)$. An easy computation shows that

$$\mathcal{E}n(r,h) = \frac{r^2}{\pi}\int_{\mathbb{R}^2} h\,.$$

If $h = \mathbb{1}_E$ is the indicator function of a set E, then $n(r,\mathbb{1}_E) = n(rE)$ is the number of zeroes in the set rE.

A usual "triad" in the asymptotic study of random variables is

VARIANCE, ASYMPTOTIC NORMALITY, LARGE FLUCTUATIONS

First, we'll discuss the variance, which is the easiest part of the triad.

1.1. The variance.

Theorem 1.1 (The variance). *For every non-zero function $h\in(L^1\cap L^2)(\mathbb{R}^2)$ and every $r>0$,*

$$\operatorname{Var} n(r,h) = r^2\int_{\mathbb{R}^2} |\widehat{h}(\lambda)|^2 M(r^{-1}\lambda)\,\mathrm{d}m(\lambda)$$

where

$$M(\lambda) = \pi^3|\lambda|^4\sum_{\alpha\geq 1}\frac{1}{\alpha^3}e^{-\frac{\pi^2}{\alpha}|\lambda|^2}\,,$$

and

$$\widehat{h}(\lambda) = \int_{\mathbb{R}^2} h(x)e^{-2\pi\mathrm{i}\,\langle\lambda,x\rangle}\,\mathrm{d}m(x)$$

is the Fourier transform of h.

This theorem was proven in [34]. The asymptotic of the variance had been known for two special cases since the work by Forrester and Honner [17]: if $h\in C_0^2$ (i.e., h is a C^2-function with compact support), then

$$\operatorname{Var} n(r,h) = \frac{\zeta(3)+o(1)}{16\pi r^2}\,\|\Delta h\|_{L^2}^2\,,\qquad r\to\infty\,, \tag{1.1}$$

while for bounded domains G with piecewise smooth boundary,

$$\operatorname{Var} n(rG) = \frac{\zeta(3/2)+o(1)}{8\pi^{3/2}}\,r\,L(\partial G)\,,\qquad r\to\infty\,. \tag{1.2}$$

Here, $\zeta(\,\cdot\,)$ is Riemann's zeta-function.

A less precise form of Theorem 1.1 might be more illustrative:

$$\operatorname{Var} n(r,h) \simeq r^{-2} \int_{|\lambda|\le r} |\widehat{h}(\lambda)|^2 |\lambda|^4 \,\mathrm{d}m(\lambda) + r^2 \int_{|\lambda|\ge r} |\widehat{h}(\lambda)|^2 \,\mathrm{d}m(\lambda)\,, \qquad (1.3)$$

where the notation $A \simeq B$ means that the quotient B/A is bounded from below and from above by positive numerical constants. The right-hand side of (1.3) interpolates $\|h\|^2_{L^2(m)}$ and $\|\Delta h\|^2_{L^2(m)}$.

1.2. Digression: "superhomogeneous" point processes.

By (1.2), the random zero process $\mathcal{Z}$ belongs to the family of translation invariant point processes with variance of the number of points in large domains proportional to the length of the boundary rather than to the area, as it would be, say, for the Poisson process. Such processes are called superhomogeneous.

A "toy model" for such processes is the point process

$$\mathcal{S} = \left\{\omega + \zeta_\omega : \omega \in \sqrt{\pi}\,\mathbb{Z}^2\right\} \qquad (1.4)$$

obtained by perturbing the lattice $\sqrt{\pi}\,\mathbb{Z}^2$ by independent standard complex Gaussian random variables ζ_ω. The normalization by $\sqrt{\pi}$ is not essential here, it is introduced to have asymptotically the same mean number of points in large areas as our process $\mathcal{Z}$ has. The choice of the square lattice is not essential either.

Curiously, the same kernel $e^{z\overline{w}}$ that occurs in the definition of random complex zeroes generates by a very different construction another interesting superhomogeneous point process $\mathcal{G}$, namely, the determinantal process whose k-point functions can be expresed in terms of the determinants formed by this kernel.

$$\rho(z_1, ..., z_k) = \pi^{-k} e^{-\sum_{i=1}^k |z_i|^2} \det\left\|e^{z_i \overline{z_j}}\right\|_{1\le i,j\le k}.$$

This process arises as the large N limit of eigenvalues of Ginibre ensemble of $N\times N$ matrices with independent standard complex Gaussian entries, and we will call it the *limiting Ginibre process*. It is known that the Ginibre point process is a special, explicitly solvable case of a one-component plasma of charged particles of one sign confined by a uniform background of the opposite sign. Though the one-component plasma has been studied by physicists for a long time, it seems that most of rigorous results still pertain only to the very special case of the Ginibre ensemble.

Resemblances and differences between the processes $\mathcal{Z}$, $\mathcal{S}$, and $\mathcal{G}$ were discussed both in the physical and the mathematical literature. For instance, the behaviour of smooth linear statistics for these three processes is quite different. In particular, decay of the variance of smooth linear statistics (1.1) distinguishes the zero process $\mathcal{Z}$ from the processes $\mathcal{G}$ and $\mathcal{S}$, since for the latter two processes, the variance of smooth linear statistics tends to the positive limit proportional to $\|\nabla h\|^2_{L^2(m)}$.

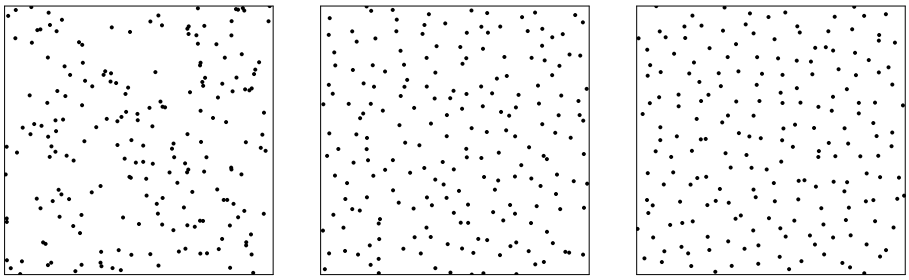


Figure 1. Samples of the Poisson process (figure by B. Virág), limiting Ginibre process, and zeroes of a GEF (figures by M. Krishnapur). Some properties of the last two processes are quite different, though the eye does not easily distinguish between them.

1.3. Asymptotic normality of fluctuations.

1.3.1. Normal fluctuations. We say that the linear statistics $n(r,h)$ have *asymptotically normal fluctuations* if the normalized linear statistics

$$\frac{n(r,h)-\mathcal{E}n(r,h)}{\sqrt{\operatorname{Var} n(r,h)}}$$

converge in distribution to the standard (real) Gaussian random variable as $r\to\infty$.

Let C_0^α, $\alpha>0$, be the class of compactly supported C^α-functions, by C_0^0 we denote the class of bounded compactly supported measurable functions.

Theorem 1.2 (Asymptotic normality). *Suppose that $h\in C_0^\alpha$ with some $\alpha\ge 0$, and that for some $\varepsilon>0$ and for every sufficiently big r, we have*

$$\operatorname{Var} n(r,h) > r^{-2\alpha+\varepsilon}\,. \tag{1.5}$$

Then the linear statistics $n(r,h)$ have asymptotically normal fluctuations.

Note that by (1.3), we always have $\operatorname{Var} n(r,h)\ge c(h)r^{-2}$ with positive $c(h)$ independent of r. Hence, for $\alpha>1$, condition (1.5) holds automatically, and we obtain the following

Corollary 1.1. *Suppose that $h\in C_0^\alpha$ with $\alpha>1$. Then the linear statistics $n(r,h)$ have asymptotically normal fluctuations.*

Using estimate (1.3), one can show that for *any bounded measurable set E of positive area*, $\operatorname{Var} n(rE)\gtrsim r$, cf. (1.2). Hence,

Corollary 1.2. *Let E be a bounded measurable set of positive area. Then the number of random complex zeroes $n(rE)$ on the set rE has asymptotically normal fluctuations.*

1.3.2. Abnormal fluctuations of linear statistics. Do there exist C_0^α-functions h with abnormal asymptotic behaviour of linear statistics $n(r,h)$? The answer is "yes", and the simplest example is provided by the function $h_\alpha = |x|^\alpha \psi(x)$, where ψ is a smooth cut-off that equals 1 in a neighborhood of the origin. Clearly, $h_\alpha \in C_0^\alpha$ and it is not difficult to show that $\operatorname{Var} n(r,h_\alpha) \simeq r^{-2\alpha}$. This shows that Theorem 1.2 is sharp on a rough power scale. The reason for the loss of asymptotic normality is that only a small neighbourhood of the origin where h_α loses its smoothness contributes to the variance of $n(r,h_\alpha)$. This neighbourhood contains a bounded number of zeroes of F, which is not consistent with the idea of normal fluctuations.

1.3.3. Comments and questions. Theorem 1.2 was preceded by a result of Sodin and Tsirelson [45, Part I]. Using the moment method and the diagrams, they showed that the fluctuations are asymptotically normal provided that $h \in C_0^2$. Their technique works in several other cases, for instance, when $h = \mathbb{1}_G$ is the indicator function of a bounded domain G with a piecewise smooth boundary. However, it seems very difficult to adapt it for proving Theorem 1.2 in its full generality.

In the case $\alpha > 0$, the proof of Theorem 1.2 is given in [34]. It uses a classical idea of S.Bernstein to approximate the random variable $n(R,h)$ by a sum of a large number of independent random variables with negligible error. Such approximation becomes possible only after we separate the high and the low frequencies in h. In this approach, independence appears as a consequence of the almost independence of the values of the G.E.F. at large distances, which we'll discuss below in Section 3.1. We do not know whether asymptotic normality holds for all functions $h \in C_0^1$, or whether the condition $r^{2\alpha} \operatorname{Var} n(r,h) \to \infty$ is already sufficient for asymptotic normality of linear statistics associated with a C_0^α-function. Also, we believe that the assertion of Theorem 1.2 can be extended to functions $h \in C^\alpha \cap L_0^2$ with $-1 < \alpha < 0$ but our current techniques seem insufficient to handle this case properly.

In the case $\alpha = 0$, the proof is given in [35]. It uses a different idea which comes from statistical mechanics. First, we show that k-point functions of the zero process $\mathcal{Z}$ are clustering, see Section 3.2 for the precise statement. Then, using clustering, we estimate the cumulants of the random variable $n(r,h)$.

It is interesting to juxtapose Theorem 1.2 with what is known for the limiting Ginibre process $\mathcal{G}$ described above. For bounded compactly supported functions h, a counterpart of Theorem 1.2 is a theorem of Soshnikov. In [47], he proved among other things that for arbitrary determinantal point processes, the fluctuations of linear statistics associated with a compactly supported bounded positive function are normal if the variance grows at least as a positive power of expectation as the intensity tends to infinity. A counterpart of the limiting case $\alpha = 2$ in Theorem 1.2 (that is, of the result from [45, Part I]) was recently found by Rider and Virág in [42]. They proved that the fluctuations for linear statistics of process $\mathcal{G}$ are normal when the test function h belongs to the

Sobolev space W^2_1. It is not clear whether there is any meaningful statement interpolating between the theorems of Soshnikov and Rider and Virág. It can happen that our Theorem 1.2 simply has no interesting counterpart for the process $\mathcal{G}$. It is also worth mentioning that the proofs in the determinantal case are quite different from ours. They are based on peculiar combinatorial identities for the cumulants of linear statistics that are a special feature of determinantal point processes.

1.4. Probability of large fluctuations. Now, we turn to the probability of exponentially rare events that, for some $r \gg 1$, $|n(r,h) - \mathcal{E}n(r,h)|$ is much bigger than $\sqrt{\mathrm{Var}(n(r,h)}$. Mostly, we consider the case when h is the indicator function of the unit disk $\mathbb{D}$; i.e., we deal with the number $n(r)$ of random zero points in the disk of large radius r centered at the origin. Recall that $\mathcal{E}n(r) = r^2$ and $\mathcal{E}\big\{(n(r) - r^2)^2\big\} \sim cr$ for $r \to \infty$ (with some $c > 0$). Hence, given $\alpha \geq \frac{1}{2}$, we need to find the order of decay of the probability $\mathcal{P}\big\{|n(r) - r^2| > r^\alpha\big\}$.

1.4.1. Naïve heuristics. The aforementioned similarity between the zero process $\mathcal{Z}$ and independent complex Gaussian perturbations $\mathcal{S}$ of the lattice $\sqrt{\pi}\,\mathbb{Z}^2$ helps to guess the correct answer.

We fix the parameter $\nu > 0$, and consider the random point set $\mathcal{S}_\nu = \{\omega + \zeta_\omega\}_{\omega \in \sqrt{\pi}\mathbb{Z}^2}$, where ζ_ω are independent, identical, radially distributed random variables with the tails $\mathcal{P}\{|\zeta_\omega| > t\}$ decaying as $\exp(-t^\nu)$ as $t \to \infty$. Set

$$n_\nu(r) = \#\{\omega \in \sqrt{\pi}\,\mathbb{Z}^2 \colon |\omega + \zeta_\omega| \leq r\}.$$

Then, for every $\alpha \geq \frac{1}{2}$ and every $\varepsilon > 0$,

$$\exp[-r^{\varphi(\alpha,\nu)+\varepsilon}] < \mathcal{P}\big\{|n_\nu(r) - r^2| > r^\alpha\big\} < \exp[-r^{\varphi(\alpha,\nu)-\varepsilon}]\,,$$

provided that r is sufficiently big. Here

$$\varphi(\alpha,\nu) = \begin{cases} 2\alpha - 1, & \frac{1}{2} \leq \alpha \leq 1; \\ (\nu+1)\alpha - \nu, & 1 \leq \alpha \leq 2; \\ (\frac{1}{2}\nu + 1)\alpha, & \alpha \geq 2\,. \end{cases}$$

Actually, one can find much sharper estimates for $\mathcal{P}\big\{|n_\nu(r) - r^2| > r^\alpha\big\}$.

This suggests that the probability $\mathcal{P}\{|n(r) - r^2| > r^\alpha\}$ we are after should decay as $\exp[-r^{\varphi(\alpha)}]$ with

$$\varphi(\alpha) = \varphi(\alpha,2) = \begin{cases} 2\alpha - 1, & \frac{1}{2} \leq \alpha \leq 1; \\ 3\alpha - 2, & 1 \leq \alpha \leq 2; \\ 2\alpha, & \alpha \geq 2\,. \end{cases}$$

1.4.2. Jancovici-Lebowitz-Manificat Law. Unfortunately, we do not know how to represent random complex zeroes as independent, or weakly correlated, Gaussian perturbations of the lattice points, so we cannot use the heuristics given above. Nevertheless, we can prove

Theorem 1.3 (JLM Law for random complex zeroes). *For every* $\alpha \geq \frac{1}{2}$ *and every* $\varepsilon > 0$,

$$\exp[-r^{\varphi(\alpha)+\varepsilon}] < \mathcal{P}\left\{|n(r) - r^2| > r^\alpha\right\} < \exp[-r^{\varphi(\alpha)-\varepsilon}]$$

for all sufficiently large $r > r_0(\alpha, \varepsilon)$ *with the same* $\varphi(\alpha)$ *as above.*

In [18], Jancovici, Lebowitz and Manificat showed that this law holds for the one-component plasma. Their derivation was not a rigorous one, except for the case of the limiting Ginibre process $\mathcal{G}$. It would be desirable to have a clear explanation why *the same* Jancovici-Lebowitz-Manificat law holds for the random processes $\mathcal{Z}$, $\mathcal{S}$, and $\mathcal{G}$ in the range $\alpha > 1$.

1.4.3. Comments and questions. The function φ from the exponent in the Jancovici-Lebowitz-Manificat Law loses smoothness at three points. Accordingly, there are three different régimes ($\frac{1}{2} < \alpha < 1$, $1 < \alpha < 2$, and $\alpha > 2$). The point $\alpha = \frac{1}{2}$ corresponds to the asymptotic normality of $n(r)$, and deviations in the range $\frac{1}{2} < \alpha < 1$ are called *moderate*. In this range, the deviation $|n(r) - r^2|$ is small compared to the length of the circumference $\{|z| = r\}$. In this case, the theorem was proven by Nazarov, Sodin, and Volberg [37]. The point $\alpha = 1$ corresponds to the classical large deviations principle. In the range $1 < \alpha < 2$, the deviation is already big compared to the length of the boundary circumference, but is still small compared to the area of the disk $\{|z| \leq r\}$. In this case, the lower bound for $\mathcal{P}\left\{|n(r) - r^2| > r^\alpha\right\}$ is due to Krishnapur [22], while the upper bound was proven in [37].

The case $\alpha = 2$ contains an estimate for the "hole probability" $\mathcal{P}\{n(r) = 0\}$. In this case, the theorem was proved by Sodin and Tsirelson [45, Part III]. A very sharp estimate of the hole probability

$$\log \mathcal{P}\{n(r) = 0\} = -\frac{3e^2}{4} r^4 + O\big(r^{\frac{18}{5}}\big), \qquad r \to \infty,$$

was recently obtained by Nishry [38]; in [39] he extended this asymptotics to a rather wide class of entire functions represented by Gaussian Taylor series. There are two interesting questions pertaining to the hole probability. We have no idea how to find the asymptotics of the expected number of random complex zeroes in the disk $R\mathbb{D}$, $R \geq r$, conditioned on the hole $\{n(r) = 0\}$. We also do not know how to extend Nishry's result from the unit disk to other bounded domains G. It seems plausible that for a large class of bounded

domains G,

$$\log \mathcal{P}\left\{n(rG)=0\right\} = -(\kappa(G)+o(1))r^4\,, \qquad r\to\infty,$$

with $\kappa(G)>0$. If this is true, how does $\kappa(G)$ depend on G?

The range $\alpha>2$ in the Jancovici-Lebowitz-Manificat Law is the "overcrowding" régime. In [22], Krishnapur proved that for $\alpha>2$,

$$\log \mathcal{P}\left\{n(r)>r^\alpha\right\} = -\big(\tfrac{1}{2}\alpha-1+o(1)\big)r^{2\alpha}\log r\,, \qquad r\to\infty\,.$$

The bounds in Theorem 1.3 are not too tight. As we've already mentioned, in some cases, much better bounds are known. It would be good to improve precision of Theorem 1.3 in other cases. For instance, to show that for $\alpha\le 2$ and for $\delta>0$ there exists the limit

$$\lim_{r\to\infty}\frac{\log \mathcal{P}\left\{|n(r)-r^2|>\delta r^\alpha\right\}}{r^{\varphi(\alpha)}}$$

and to find its value.

1.4.4. Moderate deviations for smooth linear statistics. Here is a recent result of Tsirelson [52]:

Theorem 1.4. *Let $h\in C_0^2$. Then*

$$\log \mathbb{P}\big\{rn(r,h)>t\sigma\|\Delta h\|_{L^2}\big\} = (1+o(1))\log\Big(\frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}\,\mathrm{d}x\Big)$$

and

$$\log \mathbb{P}\big\{rn(r,h)<-t\sigma\|\Delta h\|_{L^2}\big\} = (1+o(1))\log\Big(\frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}\,\mathrm{d}x\Big)\,,$$

as $r\to\infty$, $t>0$, and $t\frac{\log^2 r}{r}\to 0$. Here, $\sigma^2=\frac{\zeta(3)}{16\pi}$ (cf. (1.1)).

The proof of this theorem is quite intricate. Note that it gives bounds that are much sharper than the ones in Theorem 1.3. In the case $t=\mathrm{const}$, Theorem 1.4 gives another proof of the asymptotic normality of smooth linear statistics of random complex zeroes.

It is not clear whether the assumption $t\frac{\log^2 r}{r}\to 0$ can be replaced by a more natural one $\frac{t}{r}\to 0$. To the best of our knowledge, until now, there have been no results about large or huge deviations for smooth linear statistics of random complex zeroes when t is comparable or much larger than r.

2. Uniformity of Spreading of Random Complex Zeroes Over the Plane

Let Z be a point process in $\mathbb{R}^d$ with the distribution invariant with respect to the isometries of $\mathbb{R}^d$. A natural way to check how evenly the process Z is spread over $\mathbb{R}^d$ is to find out how far the counting measure

$$n_Z = \sum_{a \in Z} \delta_a$$

of the set Z (δ_a is the unit mass at a) is from the Lebesgue measure m_d in $\mathbb{R}^d$. We describe a convenient way to measure the distance between n_Z and m_d.

Suppose that the mean number of points of Z per unit volume equals 1. We want to partition the whole space $\mathbb{R}^d$, except possibly a subset of zero Lebesgue measure, into disjoint sets $B(a)$ of Lebesgue measure 1 indexed by sites $a \in Z$ in such a way that each set $B(a)$ is located not too far from the corresponding site $a \in Z$. In other words, we are looking for a measurable map $T\colon \mathbb{R}^d \to Z$ such that for each $a \in Z$, we have $m_d(T^{-1}\{a\}) = 1$. We also want the distances $|Tx - x|$ to be not too large. The map T is called the *transportation* (a.k.a. "matching", "allocation", "marriage", etc.) of the Lebesgue measure m_d to the set Z.

Alternatively, we can fix a lattice $\Gamma \subset \mathbb{R}^d$ with cells of unit volume, and look for a bijection $\Theta\colon \Gamma \to Z$ for which the distances $|\Theta\gamma - \gamma|$, $\gamma \in \Gamma$, are not too large. Since for each two lattices Γ_1, Γ_2 with cells of the same volume, there is a bijection $\theta\colon \Gamma_1 \to \Gamma_2$ with $\sup\{|\theta\gamma - \gamma| \colon \gamma \in \Gamma_1\} < \infty$, the choice of the lattice is not important, so we can take $\Gamma = \mathbb{Z}^d$.

Since we deal with *random* discrete sets Z, the corresponding transportation maps T (or the bijections Θ) will be *random maps*. In interesting cases (including the random complex zeroes $\mathcal{Z}$), almost surely, the transportation distances $|Tx - x|$ are unbounded, so we are interested in the rate of decay of the probability tails $\mathcal{P}\{|Tx - x| > R\}$ as $R \to \infty$.

Here we present two approaches to this problem developed in [45, Part II] and in [36]. Though we discuss only the random complex zeroes $\mathcal{Z}$, we believe that both approaches should work for other natural translation invariant point processes. At last, we recall that the random complex zero process $\mathcal{Z}$ has intensity π, not 1. For this reason, we will look for a transportation of the measure πm_2 to $\mathcal{Z}$, and for a bijection between the lattice $\sqrt{\pi}\,\mathbb{Z}^2$ and $\mathcal{Z}$.

2.1. Random complex zeroes as randomly perturbed lattice points.

Theorem 2.1 (Existence of well-localized bijection). *There exists a translation invariant random function $\xi\colon \mathbb{Z}^2 \to \mathbb{C}$ such that*

(a) the random set $\{\gamma + \xi(\gamma) \colon \gamma \in \sqrt{\pi}\,\mathbb{Z}^2\}$ is equidistributed with the random complex zeroes $\mathcal{Z}$;

(b) $\mathcal{P}\{|\xi(0)| > R\} \le \exp\left(-cR^4/\log R\right)$ for some $c > 0$ and every $R \ge 2$.

The theorem is almost optimal since the probability that the disk of radius $\lambda \geq 1$ is free of random complex zeroes is not less than $\exp\left(-C\lambda^4\right)$. It seems that the question about the existence of a matching between the lattice and $\mathcal{Z}$ with tails decaying as $\exp\left(-c\lambda^4\right)$ remains open as well as the same question for the Gaussian perturbations $\mathcal{S}$ of the lattice points and for the limiting Ginibre process $\mathcal{G}$.

It would be interesting to find a version of Theorem 2.1 with weakly correlated perturbations $\xi_{k,l}$ at large distances. This could shed some light on the reasons hidden behind the Jancovici-Lebowitz-Manificat Law.

2.1.1. Uniformly spread sequences in $\mathbb{R}^d$. The proof of Theorem 2.1 is based on a deterministic idea which might be useful in study of the uniformity of spreading of sequences and measures. We need to establish the bijection between the sets $\mathcal{Z}$ and $\sqrt{\pi}\mathbb{Z}^2$ with controlled tails of the distances $|\xi_{k,l}|$. First we look at a simpler situation when $|\xi_{k,l}|$ are uniformly bounded. It is too much to expect from a typical zero set, but let us try anyway. We say that the set $Z \in \mathbb{R}^d$ is *r-uniformly spread* over $\mathbb{R}^d$ (with density 1) if there exists a bijection between Z and a lattice with the unit volume of the cell such that the distances between Z and the corresponding lattice points do not exceed r. If such a bijection exists then clearly

$$n(U) \leq \nu(U_{+r}) \qquad \text{and} \qquad \nu(U) \leq n(U_{+r}) \tag{2.1}$$

for every $U \subset \mathbb{R}^d$; here U_{+r} stands for the r-neighbourhood of U, n is the counting measure of the set Z, and ν is the counting measure of the lattice. In fact, (2.1) is not only necessary but also *sufficient*, which is basically a well-known locally finite marriage lemma due to M. Hall and R. Rado. When verifying condition (2.1), we can replace ν by the Lebesgue measure m_d at the expense of adding a constant to r .

Now, given a locally finite measure μ on $\mathbb{R}^d$, we define $\mathrm{Di}(\mu)$ as the infimum of $r \in (0, \infty)$ such that

$$\mu(X) \leq m_d(X_{+r}) \qquad \text{and} \qquad m_d(X) \leq \mu(X_{+r})$$

for every bounded Borel set $X \subset \mathbb{R}^d$. The range of Di is $[0, +\infty]$ with the both ends included. The following theorem gives a useful upper bound for $\mathrm{Di}(\mu)$ in terms of the potential u:

Theorem 2.2 (Upper bound for the transportation distance). *Let u be a locally integrable function in $\mathbb{R}^d$ such that $\Delta u = \mu - m_d$ in the sense of distributions. Then*

$$\mathrm{Di}(\mu) \leq \mathrm{Const}_d \cdot \inf_{r>0} \left\{ r + \sqrt{\|u * \chi_r\|_\infty} \right\}.$$

Here, χ_r is the indicator function of the ball of radius r centered at the origin normalized by the condition $\|\chi_r\|_{L^1} = 1$, and $*$ denotes the convolution.

Now, we explain how Theorem 2.1 is deduced from Theorem 2.2. After smoothing, the random potential $U(z) = \log|F(z)| - \frac{1}{2}|z|^2$ is locally uniformly bounded. Still, a.s. it remains unbounded in $\mathbb{C}$, so we cannot apply Theorem 2.2 directly. The idea is to introduce on $\mathbb{C}$ a *random metric* ρ that depends on a G.E.F. F. The metric ρ is small where the random potential U is large. Then we apply a counterpart of Theorem 2.2 with the distances measured in the metric ρ, instead of the Euclidean one.

2.1.2. Comments. Theorems 2.1 and 2.2 are taken from Sodin and Tsirelson [45, Part II] (cf. [46]). In that paper, the authors proved a weaker subgaussian estimate for the tails, however, after a minor modification of the proof given therein, one gets the result formulated here. Note that the method developed in Sodin and Tsirelson [45, Part II] needs only the existence of a stationary random vector field v in $\mathbb{R}^d$ with $\operatorname{div} v = \mu - c_d m_d$. The tail estimate depends on the rate of decay of the tails of the field v or of the tails of the potential u such that $v = \nabla u$ (if such a u exists).

In the last 20 years, the concept of *uniformly spread* discrete subsets of $\mathbb{R}^d$ has appeared in very different settings. Laczkovich used uniformly spread sets in $\mathbb{R}^d$ in his celebrated solution of the Tarski's circle squaring problem [23] (see also [24]). There are various probabilistic counterparts of this notion. For instance, Ajtai, Komlós and Tusnády [1], Leighton and Shor [25], and Talagrand [51] studied a finite counterpart of this, namely, a high probability matching of a system of N^2 independent random points in the square $[0, N)^2 \subset \mathbb{R}^2$ with the grid $\mathcal{Z}^2 \cap [0, N)^2$.

2.2. Gradient transportation.

Unfortunately, the proof of Theorem 2.1 is a pure existence one. It gives us no idea about what the (almost) optimal transportation of the Lebesgue measure to the zero process $\mathcal{Z}$ looks like. Now, we discuss another approach, namely, the transportation by the gradient flow of a random potential. The main advantage of this approach is that it yields a quite natural and explicit construction for the desired transportation.

2.2.1. Basins of zeroes. Let $U(z) = \log|F(z)| - \frac{1}{2}|z|^2$ be the random potential corresponding to the G.E.F. F. It is easy to check that the distribution of U is invariant with respect to the isometries of the plane. We shall call any integral curve of the differential equation

$$\frac{dZ}{dt} = -\nabla U(Z)$$

a *gradient curve* of the potential U. We orient the gradient curves in the direction of decrease of U (this is the reason for our choice of the minus sign in the differential equation above). If $z \notin \mathcal{Z}$, and $\nabla U(z) \neq 0$, by Γ_z we denote the (unique) gradient curve that passes through the point z.

Definition 2.1. Let a be a zero of the G.E.F. F. The *basin* of a is the set

$$B(a) = \{z \in \mathbb{C} \setminus \mathcal{Z} \colon \nabla U(z) \neq 0, \text{ and } \Gamma_z \text{ terminates at } a\}.$$

Clearly, each basin $B(a)$ is a connected open set, and $B(a') \cap B(a'') = \varnothing$ if a' and a'' are two different zeroes of F. Remarkably, *all bounded basins have the same area* π. Indeed, $\frac{\partial U}{\partial n} = 0$ on $\partial B(a)$ and therefore, applying the Green formula and recalling that the distributional Laplacian of U equals $\Delta U = 2\pi \sum_{a \in \mathcal{Z}_F} \delta_a - 2m$, one gets

$$1 - \frac{mB(a)}{\pi} = \frac{1}{2\pi} \iint_{B(a)} \Delta U(z)\, \mathrm{d}m(z) = \frac{1}{2\pi} \int_{\partial B(a)} \frac{\partial U}{\partial n}(z)\, |dz| = 0\,;$$

i.e., $mB(a) = \pi$. The picture below helps to visualize what's going on.

Figure 2. Random partition of the plane into domains of equal area generated by the gradient flow of the random potential U (figure by M. Krishnapur). The lines are gradient curves of U, the black dots are random zeroes. Many basins meet at the same local maximum, so that two of them meet tangentially, while the others approach it cuspidally forming long, thin tentacles.

2.2.2. Results.

Theorem 2.3 (Random partition). *Almost surely, each basin is bounded by finitely many smooth gradient curves (and, thereby, has area* π*), and*

$$\mathbb{C} = \bigcup_{a \in \mathcal{Z}} B(a)$$

up to a set of measure 0 *(more precisely, up to countably many smooth boundary curves).*

The tails of this random partition have three characteristic exponents **1**, $\frac{\mathbf{8}}{\mathbf{5}}$, and **4**. The probability that the diameter of a particular basin is greater than R is exponentially small in R. Curiously enough, the probability that a given point z lies at a distance larger than R from the zero of F it is attracted to decays much faster: as $e^{-R^{8/5}}$. This is related to long thin tentacles seen on the picture around some basins. They increase the typical diameter of the basins though the probability that a given point z lies in such a tentacle is very small. At last, given $\varepsilon > 0$, the probability that it is impossible to throw away $\varepsilon\%$ of the area of the basin so that the diameter of the remaining part is less than R decays as e^{-R^4}. All three exponents are optimal. The proofs of these results rely on the following long gradient curve theorem.

Theorem 2.4 (Long gradient curve). *Let $R \geq 1$. Let $Q(R)$ be the square centered at the origin with side length R. The probability of the event that there exists a gradient curve joining $\partial Q(R)$ with $\partial Q(2R)$ does not exceed $Ce^{-cR(\log R)^{3/2}}$.*

The proof of this theorem is, unfortunately, rather long and complicated. It might be helpful for the reader to look at the first version of [36] posted in the `arxiv` where the authors gave a more transparent proof of a weaker upper bound $Ce^{-cR\sqrt{\log R}}$ in the long gradient curve theorem.

2.2.3. Comments and questions. Gradient transportation was introduced by Sodin and Tsirelson [45, Part II] and studied by Nazarov, Sodin, Volberg in [36].

There are several questions related to the statistics of our random partition of the plane. It is not difficult to show that, almost surely, any given point $z \in \mathbb{C}$ belongs to some basin. We denote that basin by B_z, and the corresponding sink by a_z. We say that two basins are neighbours if they have a common gradient curve on the boundary. By N_z we denote the number of basins B neighbouring the basin B_z. Clearly, N_z equals the number of saddle points of the potential U connected with the sink a_z by gradient curves. Heuristically, since almost surely each saddle point is connected with two sinks,

$$\mathcal{E}N_z = 2\,\frac{\text{mean number of saddle points per unit area}}{\text{mean number of zeroes per unit area}}\,.$$

Douglas, Shiffman and Zelditch proved in [15] that the mean number of saddle points of U per unit area is $\frac{4}{3\pi}$. (They proved this for another closely related "elliptic model" of Gaussian polynomials. It seems that their proof also works for G.E.F.'s) This suggests that $\mathcal{E}N_z = \frac{8}{3}$. Another characteristic of the random partition is the number of basins that meet at the same local maximum. Taking into account the result from [15], we expect that its average equals 8.

The next question concerns the topology of our random partition of the plane. By the *skeleton* of the gradient flow we mean the connected planar graph

with vertices at local maxima of U and edges corresponding to the boundary curves of the basins. The graph may have multiple edges and loops. We do not know whether there are any non-trivial topological restrictions on finite parts of the skeleton that hold almost surely.

In [11] Chatterjee, Peled, Peres, Romik applied the ideas from [36] to study the gradient transportation of the Lebesgue measure to the Poisson point process in $\mathbb{R}^d$ with $d \geq 3$ (they called it 'gravitational allocation'). Their work required a delicate and thorough analysis of the behaviour of the Newtonian potential of the Poisson point process. It's worth mentioning that a very different construction of the *stable marriage* between the Lebesgue measure m_d and the Poisson process in $\mathbb{R}^d$ with $d \geq 2$ was developed by Hoffman, Holroyd and Peres in [20]. The case $d = 2$ is especially interesting: see the recent work by Holroyd, Pemantle, Peres, Schramm [19].

3. Almost Independence and Correlations

3.1. Almost independence at large distances. The covariance function of the normalized Gaussian process $F^*(z) = F(z)e^{-\frac{1}{2}|z|^2}$ equals

$$e^{z\overline{w}-\frac{1}{2}|z|^2-\frac{1}{2}|w|^2} = e^{\mathrm{iIm}(z\overline{w})-\frac{1}{2}|z-w|^2},$$

which decays very fast as $|z-w|$ grows. This suggests an idea that the zeroes of G.E.F.'s must be "almost independent" on large distances. Still the precise formulation of this independence property is not obvious: due to analyticity of F, if we know the process F^* in a neighbourhood of some point, we know it everywhere on the plane.

It is not difficult to show that two standard complex Gaussian random variables with small covariance can be represented as small perturbation of two *independent* standard complex Gaussian random variables. Developing this idea, we show that if $\{K_j\}$ is a collection of well-separated compact sets, then the restrictions $F^*\big|_{K_j}$ of normalized process F^* can be simultaneously approximated by restrictions $F_j^*\big|_{K_j}$ of normalized *independent* realizations of G.E.F.'s F_j with high precision and the probability very close to 1. This is a very useful principle that lies in the core of the proofs of most of the results described above. Here is the precise statement [34]:

Theorem 3.1 (Almost independence). *Let F be a G.E.F.. There exists a numerical constant $A > 1$ with the following property. Given a family of compact sets K_j in $\mathbb{C}$ with diameters $d(K_j)$, let $\lambda_j \geq \max\{d(K_j), 2\}$. Suppose that $A\sqrt{\log \lambda_j}$-neighbourhoods of the sets K_j are pairwise disjoint. Then*

$$F^* = F_j^* + G_j^* \qquad \text{on } K_j,$$

where F_j are independent G.E.F.'s and for every j, we have

$$\mathcal{P}\left\{\max_{K_j}|G_j^*|\geq\lambda_j^{-1}\right\}\lesssim e^{-\lambda_j}\,.$$

Less general versions of this result were proven in [36, 37].

The proof of Theorem 3.1 goes as follows. First, for each compact set K_j, we choose a sufficiently dense net Z_j and consider the bunch $N_j=\{v_z\colon z\in Z_j\}$ of unit vectors $v_z=F^*(z)$ in the Hilbert space of complex Gaussian random variables. Since the compact sets K_j are well-separated, the bunches N_j are almost orthogonal to each other. Then we slightly perturb the vectors v_z without changing the angles between the vectors within each bunch N_j, making the bunches orthogonal to each other. More accurately, we construct new bunches $\widetilde{N}_j=\{\widetilde{v}_z\colon z\in Z_j\}$ so that for $z\in Z_j$, $\zeta\in Z_k$,

$$\langle\widetilde{v}_z,\widetilde{v}_\zeta\rangle=\begin{cases}\langle v_z,v_\zeta\rangle & \text{for } j=k,\\ 0 & \text{for } j\neq k\end{cases}$$

with good control of the errors $\|v_z-\widetilde{v}_z\|$. Then we extend the Gaussian bunches $\{\widetilde{v}_z e^{\frac{1}{2}|z|^2}\colon z\in Z_j\}$ to *independent* G.E.F.'s F_j. The difference $G_j=F-F_j$ is a random entire function that is small on the net Z_j with probability very close to one. At the last step of the proof, using some simple complex analysis, we show that G_j^* is small everywhere on K_j.

3.2. Uniform estimates of k-point functions. Clustering.

There is yet another way (originated in statistical mechanics) to describe point processes by the properties of their k-point correlation functions. Recall that the k-point function $\rho=\rho_k$ of the zero process $\mathcal{Z}$ is a symmetric function on $\mathbb{C}^k$ defined outside of the diagonal subset

$$\mathtt{Diag}(\mathbb{C}^k)=\{(z_1,\dots,z_k)\colon z_i=z_j \text{ for some } i\neq j\}$$

by the formula

$$\rho(z_1,\dots,z_k)=\lim_{\varepsilon\to 0}\frac{p_\varepsilon(z_1,\dots,z_k)}{(\pi\varepsilon^2)^k} \tag{3.1}$$

where $p_\varepsilon(z_1,\dots,z_k)$ is the probability that each disk $\{|z-z_j|\leq\varepsilon\}$, $1\leq j\leq k$, contains at least one point of $\mathcal{Z}$. The k-point functions describe correlations within k-point subsets of the point process. Estimates for the k-point functions are crucial for understanding many properties of point processes. The following results taken from [35] provide rather complete quantitative information about the behaviour of the k-point functions of random complex zeroes.

The first result treats the local behaviour of k-point functions. It appears that for a wide class of non-degenerate Gaussian analytic functions, the k-point functions of their zeroes exhibit universal local repulsion when some of the variables $z_1,\dots,z_k$ approach each other.

Recall that a Gaussian analytic function $f(z)$ in a plane domain $G \subseteq \mathbb{C}$ is the sum

$$f(z) = \sum_n \zeta_n f_n(z)$$

of analytic functions $f_n(z)$ such that

$$\sum_n |f_n(z)|^2 < \infty \quad \text{locally uniformly on } G,$$

where ζ_n are independent standard complex Gaussian coefficients. By $\rho_f = \rho_f(z_1, ..., z_k)$ we denote the k-point function of the zero set of the function f. It is a symmetric function defined outside the diagonal set $\mathtt{Diag}(G^k)$ as in (3.1).

We skip the technical definition of d-degeneracy, which we use in the assumptions of the next theorem, and only mention that Gaussian Taylor series (either infinite, or finite)

$$f(z) = \sum_{n \geq 0} \zeta_n c_n z^n$$

are d-nondegenerate, provided that $c_0, c_1, ..., c_{d-1} \neq 0$. In particular, the G.E.F. is d-nondegenerate for every positive integer d.

Theorem 3.2 (Local universality of repulsion). *Let f be a $2k$-nondegenerate Gaussian analytic function in a domain G, let ρ_f be a k-point function of zeroes of f, and let $K \subset G$ be a compact set. Then there exists a positive constant $C = C(k, f, K)$ such that, for any configuration of pairwise distinct points $z_1, ..., z_k \in K$,*

$$C^{-1} \prod_{i<j} |z_i - z_j|^2 \leq \rho_f(z_1, ..., z_k) \leq C \prod_{i<j} |z_i - z_j|^2 .$$

The next result is a clustering property of zeroes of G.E.F.'s. It says that if the variables in $\mathbb{C}^k$ can be split into two groups located far from each other, then the function ρ_k almost equals the product of the corresponding factors. This property is another manifestation of almost independence of points of the process at large distances. It plays a central rôle in the proof of the asymptotic normality theorem 1.2 for bounded measurable functions.

For a non-empty subset $I = \{i_1, ..., i_\ell\} \subset \{1, 2, ..., k\}$, we set $Z_I = \{z_{i_1}, ..., z_{i_\ell}\}$. We denote by

$$d(Z_I, Z_J) = \inf_{i \in I, j \in J} |z_i - z_j|$$

the distance between the configurations Z_I and Z_J.

Theorem 3.3 (Clustering property). *For each $k \geq 2$, there exist positive constants C_k and Δ_k such that for each configuration Z of size k and each partition of the set of indices $\{1, 2, ..., k\}$ into two non-empty subsets I and J with $d(Z_I, Z_J) \geq 2\Delta_k$, one has*

$$1 - \varepsilon \leq \frac{\rho(Z)}{\rho(Z_I)\rho(Z_J)} \leq 1 + \varepsilon \quad \text{with} \quad \varepsilon = C_k e^{-\frac{1}{2}(d(Z_I, Z_J) - \Delta_k)^2} . \tag{3.2}$$

Combining Theorems 3.2 and 3.3, and taking into account the translation invariance of the point process $\mathcal{Z}$, we obtain a uniform estimate for ρ_k valid in the whole $\mathbb{C}^k$:

Theorem 3.4. *For each $k \geq 1$, there exists a positive constant C_k such that for each configuration $(z_1, ..., z_k)$,*

$$C_k^{-1} \prod_{i<j} \ell(|z_i - z_j|) \leq \rho(z_1, ..., z_k) \leq C_k \prod_{i<j} \ell(|z_i - z_j|)\,,$$

where $\ell(t) = \min(t^2, 1)$.

The proofs of Theorems 3.2 and 3.3 start with the classical Kac-Rice-Hammersley formula [21, Chapter 3]:

$$\rho_f(z_1, ..., z_k) = \int_{\mathbb{C}^k} |\eta_1|^2 ... |\eta_k|^2\, \mathcal{D}_f(\eta'; z_1, ..., z_k)\, \mathrm{d}m(\eta_1) ... \mathrm{d}m(\eta_k), \tag{3.3}$$

where $\mathcal{D}_f(\,\cdot\,; z_1, ..., z_k)$ is the density of the joint probability distribution of the random variables

$$f(z_1),\ f'(z_1),\ \dots,\ f(z_k),\ f'(z_k)\,, \tag{3.4}$$

and $\eta' = (0, \eta_1, ..., 0, \eta_k)^{\mathrm{T}}$ is a vector in $\mathbb{C}^{2k}$. Since the random variables (3.4) are complex Gaussian, one can rewrite the right-hand side of (3.3) in a more explicit form

$$\rho_f(z_1, ..., z_k) = \frac{1}{\pi^{2k} \det \Gamma_f} \int_{\mathbb{C}^k} |\eta_1|^2 ... |\eta_k|^2 e^{-\frac{1}{2}\langle \Gamma_f^{-1} \eta', \eta' \rangle} \mathrm{d}m(\eta_1) ... \mathrm{d}m(\eta_k), \tag{3.5}$$

where $\Gamma_f = \Gamma_f(z_1, ..., z_k)$ is the covariance matrix of the random variables (3.4). We consider the linear functionals

$$Lf = \sum_{j=1}^{k} [\alpha_j f(z_j) + \beta_j f'(z_j)] = \frac{1}{2\pi \mathrm{i}} \int_\gamma f(z) r^L(z)\, \mathrm{d}z,$$

where

$$r^L(z) = \sum_{j=1}^{k} \left[\frac{\alpha_j}{z - z_j} + \frac{\beta_j}{(z - z_j)^2} \right],$$

and $\gamma \subset K$ is a smooth contour that bounds a domain $G' \subset K$ that contains the points $z_1, ..., z_k$. Observe that for every vector $\delta = (\alpha_1, \beta_1, ..., \alpha_k, \beta_k)^{\mathrm{T}}$ in $\mathbb{C}^{2k}$, we have

$$\langle \Gamma_f \delta, \delta \rangle = \mathcal{E} |Lf|^2\,.$$

This observation allows us to estimate the matrix Γ_f^{-1}, and hence the integral on the right-hand side of (3.5), using some simple tools from the theory of analytic functions of one complex variable.

We note that using another approach to analyzing the right-hand side of (3.5), Bleher, Shiffman, and Zelditch proved in [5] that if the points z_i are well separated from each other, i.e.,

$$\min_{i\neq j} |z_i - z_j| \geq \delta > 0,$$

then some estimate similar to (3.2) holds with a factor $C(k,\delta)$ instead of C_k. Unfortunately, in this form the result is difficult to apply. For instance, it does not yield the boundedness of the k-point functions on the whole $\mathbb{C}^k$, and we could not use it for the proof of Theorem 1.2. On the other hand, the result of Bleher, Shiffman and Zelditch is valid for a wider class of zero point processes.

Part II. Random Nodal Lines

4. Gaussian Spherical Harmonic and Gaussian Plane Wave

We introduce two remarkable Gaussian random functions closely related to each other: the *Gaussian spherical harmonic* on the two-dimensional sphere $\mathbb{S}^2$ and its scaling limit, the *Gaussian plane wave*. The study of random plane waves, and in particular, of their nodal portraits, originated in applied mathematics and goes back to M. S. Longuet-Higgins [27] who computed various statistics of nodal lines for Gaussian random waves in connection with the analysis of ocean waves. One of the reasons for the recent interest in random plane waves is the heuristic principle proposed by M. V. Berry [3] called 'the random wave conjecture'. This principle says that the behaviour of high-energy Laplace eigenfunctions in the case when the corresponding geodesic flow is ergodic (the so called 'highly excited quantum chaotic eigenfunctions') should resemble the behaviour of Gaussian random waves. More generally, one would expect that the random spherical harmonic can serve as a good model for the typical behaviour of high-energy Laplace eigenfunctions on a compact surface endowed with a smooth Riemannian metric.

4.1. Spherical harmonics. The spherical harmonic of degree n is a real-valued eigenfunction of the Laplacian (with the minus sign) on the two-dimensional sphere $\mathbb{S}^2$ corresponding to the eigenvalue $\lambda_n = n(n+1)$. Equivalently, it is a trace of a homogeneous harmonic polynomial in $\mathbb{R}^3$ of degree n on the unit sphere. Let $\mathcal{H}_n$ be the $2n+1$-dimensional real Hilbert space of spherical harmonics of degree n equipped with the $L^2(\mathbb{S}^2)$-norm. The *Gaussian spherical harmonic* f is the sum

$$f_n = \sum_{k=-n}^{n} \xi_k Y_k$$

where ξ_k are independent identically distributed mean zero Gaussian (real) random variables with $\mathcal{E}\xi_k^2 = \frac{1}{2n+1}$ and $\{Y_k\}$ is an orthonormal basis of $\mathcal{H}_n$, so $\mathcal{E}\|f\|^2_{L^2(\mathbb{S}^2)} = 1$. As a random function, f_n does not depend on the choice of the basis $\{Y_k\}$ in $\mathcal{H}_n$. Since the scalar product in the Hilbert space $\mathcal{H}_n$ is invariant under rotations of the unit sphere, the distribution of the random spherical harmonic f_n is also rotation invariant. The covariance function of the Gaussian spherical harmonic equals

$$\mathcal{E}\{f_n(x)f_n(y)\} = P_n(\cos\Theta(x,y))$$

where $\Theta(x,y)$ is the angle between x and y, and P_n is the Legendre polynomial of degree n normalized by $P_n(1) = 1$.

4.2. Random plane waves. Now, we turn to the Gaussian plane wave. Informally speaking, it is the two-dimensional Fourier transform of the white noise on the unit circumference $\mathbb{S}^1 \subset \mathbb{R}^2$. More formally, we start with the Hilbert space $L^2_{\mathtt{sym}}(\mathbb{S}^1)$ that consists of complex valued L^2-functions φ on $\mathbb{S}^1$ satisfying the symmetry condition

$$\varphi(-\lambda) = \overline{\varphi(\lambda)}, \quad \lambda \in \mathbb{S}^1,$$

and consider the Fourier image of this space $\mathcal{H} = \mathcal{F}L^2_{\mathtt{sym}}(\mathbb{S}^1)$ with the scalar product inherited from $L^2_{\mathtt{sym}}(\mathbb{S}^1)$. The space $\mathcal{H}$ consists of real-analytic functions

$$\Phi(x) = \int_{\mathbb{S}^1} e^{\mathrm{i}x\cdot\lambda}\varphi(\lambda)\,\mathrm{d}m(\lambda)$$

(m is the Lebesgue measure on $\mathbb{S}^1$) satisfying the Helmholtz equation $\Delta\Phi + \Phi = 0$. The Gaussian plane wave is the sum of the random series

$$F = \sum_k \eta_k \Phi_k$$

where η_k are standard identically distributed independent (real) Gaussian random variables, and $\{\Phi_k\}$ is an orthonormal basis in $\mathcal{H}$. The series converges almost surely, and its sum is again a real analytic function in $\mathbb{R}^2$ satisfying the same Helmholtz equation. This construction does not depend on the choice of the basis $\{\Phi_k\}$, and the distribution of the random function F is invariant with respect to translations and rotations of the plane (since the norm in $\mathcal{H}$ is translation and rotation invariant).

Applying the Fourier transform to the standard orthonormal basis $\{\lambda^m\}_{m\in\mathbb{Z}}$ in $L^2(\mathbb{S}^1)$, we get the functions $\mathrm{i}^m J_m(r)e^{\mathrm{i}m\theta}$ where (r,θ) are polar coordinates, and J_m is the Bessel function of order m. This yields a more explicit formula for the Gaussian plane wave:

$$F(x) = \mathrm{Re}\sum_{m\in\mathbb{Z}} \zeta_m J_{|m|}(r)e^{\mathrm{i}m\theta}, \qquad x = (r,\theta),$$

where ζ_m are independent identically distributed complex Gaussian random variables with $\mathcal{E}|\zeta_m|^2 = 2$.

The covariance function of F (which is the same as the reproducing kernel of the space $\mathcal{H}$) is given by the Bessel kernel:

$$\mathcal{E}\big\{F(x)F(y)\big\} = J_0(|x-y|)\,.$$

It is worth mentioning that there are other constructions of random plane 'monochromatic' waves as random linear combinations ('superpositions') of elementary plane waves $e_\lambda(x) = e^{\mathrm{i}\lambda\cdot x}$. For instance, following Oravecz, Rudnick, Wigman [40] and Rudnick, Wigman [43], one can consider 'arithmetic random waves'

$$h_N(x) = \operatorname{Re}\sum_\nu \zeta_\nu e^{2\pi\mathrm{i}(\nu\cdot x)}$$

where ζ_ν are independent identically distributed complex Gaussian random variables with $\mathcal{E}|\zeta_m|^2 = 2$, and the sum is taken over $\nu \in \mathbb{Z}^2$ with $|\nu|^2 = N$. This model remarkably combines analysis and probability theory with the number theory. Its covariance function

$$\mathcal{E}\big\{h_N(x)h_N(y)\big\} = \sum_\nu \cos 2\pi\big(\nu\cdot(x-y)\big)$$

has a more erratic behaviour than the covariance functions of the Gaussian spherical harmonic and the Gaussian plane wave.

4.3. Random plane waves as scaling limits of random spherical harmonics. The Gaussian plane wave F is *a scaling limit* of the Gaussian spherical harmonic f_n when $n \to \infty$. This is a very special case of a result of Zelditch [55] pertaining to a wide class of Riemannian smooth surfaces, in particular, to all real-analytic Riemannian surfaces.

Informally, for any fixed R, the restrictions of the Gaussian functions f_n on spherical disks of radius R/n converge as random processes to the restriction of F on the euclidean disk of radius R. More formally, we fix a point $x_0 \in \mathbb{S}^2$, and define the random Gaussian function F_n on the tangent plane $T_{x_0}\mathbb{S}^2$ by

$$F_n(u) = \big(f_n \circ \exp_{x_0}\big)\left(\tfrac{u}{n}\right), \tag{4.1}$$

where $\exp_{x_0} : T_{x_0}\mathbb{S}^2 \to \mathbb{S}^2$ is the exponential map. After this scaling, the covariance equals

$$\mathcal{E}\big\{F_n(u)F_n(v)\big\} = P_n\left(\cos\Theta\left(\exp_{x_0}\left(\tfrac{u}{n}\right), \exp_{x_0}\left(\tfrac{v}{n}\right)\right)\right)$$

When n goes to ∞, the angle between the points $\exp_{x_0}\left(\frac{u}{n}\right)$, and $\exp_{x_0}\left(\frac{v}{n}\right)$ on the sphere is equivalent to $|u-v|/n$ (locally uniformly in u and v). Then by classical Hilb's asymptotics of the Legendre polynomials [50, Theorem 8.21.6], the scaled covariance function $\mathcal{E}\big\{F_n(u)F_n(v)\big\}$ converges to the Bessel kernel $J_0(|u-v|)$ locally uniformly in u and v.

5. Nodal Portrait

In most cases, the basic questions about the asymptotic behaviour of the nodal portrait of the Gaussian spherical harmonic f_n as $n \to \infty$, and their counterparts for the Gaussian plane wave in the 'large area limit' are equivalent to each other. In what follows, we concentrate on spherical harmonic versions which are somewhat easier to formulate.

For the spherical harmonic $g \in \mathcal{H}_n$, we denote by $Z(g) = \{x \in \mathbb{S}^2 \colon g(x) = 0\}$ its nodal set. The connected components of the complement $\mathbb{S}^2 \setminus Z(g)$ are called nodal domains of g. The following (deterministic) facts are special cases of well-known results valid for Laplace eigenfunctions on smooth Riemannian surfaces:

Theorem 5.1. *There is a positive numerical constants C such that for each $g \in \mathcal{H}_n$, the nodal set $Z(g)$ is a Cn^{-1}-net on $\mathbb{S}^2$.*

Theorem 5.2. *There is a positive numerical constant $c > 0$ such that for each $g \in \mathcal{H}_n$, every nodal domain of g contains a disk of radius cn^{-1}.*

Together with Figure 3, this gives a very rough idea of how the nodal portraits of a spherical harmonic of large degree should look.

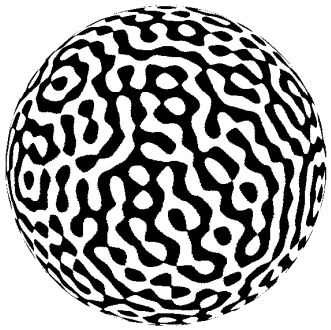


Figure 3. Nodal portrait of the Gaussian spherical harmonic of degree 40 (figure by A. Barnett)

One can find more information about the geometry and the topology of the nodal portraits of spherical harmonics (and more generally, of high-energy Laplace eigenfunctions on smooth Riemannian surfaces) in the pioneering works of Donnelly and Fefferman [12, 13, 14], as well as in the more recent works of Eremenko, Jackobson, and Nadirashvili [16], Mangoubi [29], Nazarov, Polterovich and Sodin [31], and Zelditch [55]. Still, our understanding of nodal portraits is rather restricted, and, in our opinion, this classical area of analysis is very much underdeveloped.

5.1. Length of the nodal set. The basic characteristics of the nodal set of a spherical harmonic g are its length $L(g)$ and the number $N(g)$ of

connected components (which is one less than the number of nodal domains). Useful classical integral formulas for the length due to Poincaré and to Kac and Rice make the length a somewhat easier object for a study. For instance, one can prove

Theorem 5.3. *There exists a positive numerical constant C such that for each $g \in \mathcal{H}_n$, $C^{-1}n \le L(g) \le Cn$*

This is a special case of a more general result valid for Laplace eigenfunctions corresponding to large eigenvalues (with n replaced by $\sqrt{\lambda}$). The lower bound is valid for any smooth Riemannian surface (this is a result of Brüning [10]), while the upper bound was proven by Donnelly and Fefferman [12] for real-analytic surfaces. In the smooth category, it was conjectured by S. T. Yau, and still remains open in spite of many efforts. Note that one can easily deduce the upper bound in Theorem 5.3 from the fact that spherical harmonics are restrictions of polynomials (that is, without using the deep result of Donnelly and Fefferman).

For the Gaussian spherical harmonic, Bérard showed in [2] that

Theorem 5.4. $\mathcal{E}L(f_n) = \pi\sqrt{2\lambda_n} = \sqrt{2}\,\pi n + O(1)$.

The question about the variance is more delicate. Recently, Wigman [53] confirmed a guess made by M. V. Berry [4] in a slightly different context:

Theorem 5.5. *For $n \to \infty$,*

$$\text{variance of } L(f_n) = \frac{65}{32}\log n + O(1)\,.$$

The proof of this theorem is based on a very careful analysis of asymptotic cancelations that appear in the Kac-Rice integral representation of the variance of $L(f_n)$.

5.2. The number of connected components. There are few classical facts about the number of components $N(g)$. The celebrated Courant nodal domain theorem yields

Theorem 5.6. *For every $g \in \mathcal{H}_n$, $N(g) \le n^2$.*

For large n, this upper bound was improved by Pleijel [41] to $0.69n^2$. Apparently, the sharp asymptotic upper bound is not known yet. Simple examples show that it cannot be less than $(\frac{1}{2} + o(1))n^2$. H. Lewy [26] gave an elegant construction of spherical harmonics of any degree n whose nodal sets have one component for odd n and two components for even n, which proves that no non-trivial lower bound for $N(g)$ is possible.

Till recently, nothing had been known about the asymptotic properties of the random variable $N(f_n)$ when the degree n is large. The principal difficulty

is its *non-locality*: observing the nodal curves only locally, one cannot make any definite conclusion about the number of connected components. Several years ago Blum, Gnutzmann, and Smilansky [6] raised a question about the distribution of the number of nodal domains of high-energy Laplace eigenfunctions. In the ergodic case, in accordance with Berry's heuristic principle, they suggested to find this distribution for Gaussian random plane waves and performed the corresponding numerics. To compute this distribution, Bogomolny and Schmit proposed in [8] an elegant percolation-like lattice model for description of nodal domains of random Gaussian plane waves. This model completely ignores the (quite big) correlations between the values of the random function f_n at different points but nevertheless agrees with numerics pretty well. This agreement is probably due to some hidden 'universality law' rather then the possibility to directly reduce one model to another.

5.3. Bogomolny-Schmit percolation-like model.

The Bogomolny-Schmit *hypothesis* is that the distribution of nodal domains $N(f_n)$ is roughly the same as in the following critical percolation model. Consider the square lattice with the total number of sites equal to $\left(\mathcal{E}L(f_n)\right)^2$, that is proportional to n^2, and change at each site the line crossing to one of the two equiprobable avoided crossing, as shown in the following figure. At different

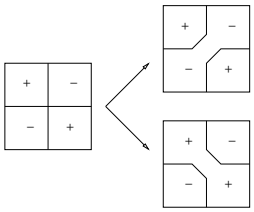


Figure 4. Avoided nodal crossings in the Bogomolny-Schmit model

sites, the changes are independent.

Then Bogomolny and Schmit introduce two dual square lattices: the 'blue one' with vertices at the cells of the grid where the function is positive, and the 'red one' with vertices at the cells of the grid where the function is negative. Each realization of the random process generates two graphs, the blue one whose vertices are the blue lattice points and the red one whose vertices are the red lattice points. Two vertices are connected by an edge if the corresponding cells of the grid belong to the same nodal domain of the random function. Each of these graphs uniquely determines the whole picture, so it suffices to consider only one of them, and each of them represents the critical bond percolation on the corresponding square lattice. Then using some heuristics coming from statistical mechanics, Bogomolny and Schmit predicted that for $n \to \infty$,

$$\mathcal{E}N(f_n) = (a + o(1))n^2,$$

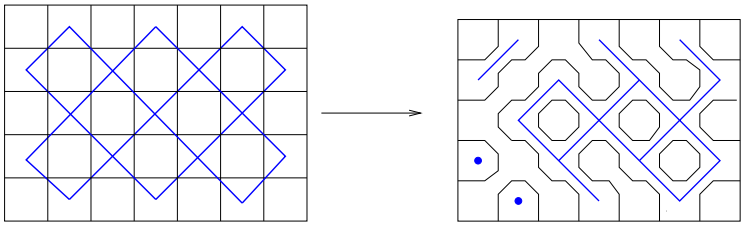


Figure 5. Bond percolation on the 'blue' lattice

and

$$\text{variance of } N(f_n) = (b + o(1))n^2,$$

with explicitly computed positive numerical constants a and b. They also argued that the fluctuations of the random variable $N(f_n)$ are asymptotically Gaussian when $n \to \infty$, and concluded their work with a remarkable prediction of the power distribution law for the areas of nodal domains, based on the percolation theory.

It would be interesting to test numerically whether the Bogomolny-Schmit model persists for random linear combinations of plane waves $e^{ik\cdot x}$ with *different* wave numbers k.

5.4. Rigorous results. Recently, we showed in [32] that, in accordance with one of the Bogomolny and Schmit predictions, $\mathcal{E}N(f)/n^2$ tends to a positive limit when $n \to \infty$, though our proof does not provide us with an explicit value of the limit a, so we cannot juxtapose it with the one predicted by Bogomolny and Schmit. In addition, we proved that the random variable $N(f)/n^2$ concentrates around this limit exponentially. Since for any spherical harmonic $g \in \mathcal{H}$, the total length of its nodal set $Z(g)$ does not exceed $\mathrm{Const}\, n$, our result yields that, for a typical spherical harmonic, most of its nodal domains have diameters comparable to $1/n$.

Theorem 5.7 (Number of nodal domains). *There exists a constant $a > 0$ such that, for every $\varepsilon > 0$, we have*

$$\mathcal{P}\left\{\left|\frac{N(f_n)}{n^2} - a\right| > \varepsilon\right\} \le C(\varepsilon)e^{-c(\varepsilon)n}$$

where $c(\varepsilon)$ and $C(\varepsilon)$ are some positive constants depending on ε only.

The exponential decay in n in Theorem 5.7 cannot be improved: we showed that given a positive and arbitrarily small κ,

$$\mathcal{P}\left\{N(f_n) < \kappa n^2\right\} \ge e^{-C(\kappa)n}.$$

On the other hand, our proof of Theorem 5.7 gives a very small value $c(\varepsilon) \simeq \varepsilon^{15}$ and it would be nice to reduce the power 15 of ε to something more reasonable. The question about the variance of $N(f_n)$ remains open.

The last but not least remark is that the proof of Theorem 5.7 uses only relatively simple tools from the classical analysis, which we believe may work in a more general setting of random functions of several real variables (and for higher Betti numbers), while it seems that the Bogomolny-Schmit model is essentially a two-dimensional one.

5.5. Related work. We are aware of several encouraging attempts to tackle similar questions in different contexts. In [49] (motivated by some engineering problems), Swerling estimated from below and from above the mean number of connected components of the *level lines* $Z(t, f) = \{f = t\}$ of a random Gaussian trigonometric polynomial f of two variables of a given degree n. His method is based on estimates of the integral curvature of the level line $Z(t, f)$. The estimates are rather good when the level t is separated from zero, but as $t \to 0$ they are getting worse and, unfortunately, give nothing when $t = 0$.

In the paper [28], Malevich considered C^2-smooth Gaussian random functions f on $\mathbb{R}^2$ with positive covariance function that decays polynomially as the distance between the points tends to infinity. She proved that for $T \geq T_0$,

$$C^{-1}T^2 \leq \mathcal{E}N(T) \leq CT^2\,,$$

where $N(T)$ is the number of the connected components of the zero set of f that are contained in the square $[0, T] \times [0, T]$, and C is a positive numerical constant. Her proof relies heavily on the positivity property of the covariance function that does not hold for Gaussian spherical harmonics or for Gaussian trigonometric polynomials.

In the recent paper [30], Mischaikow and Wanner studied the following question. Suppose f is a random smooth function on the square $[0, 1]^2$ with periodic boundary conditions and that the signs of f are computed at the vertices of the grid with mesh δ. *How small must δ be (in terms of the a priori smoothness constants of f) in order to recover the Betti numbers of the sets $\{f > 0\}$ and $\{f < 0\}$ with probability close to one ?* In particular, they show that for random trigonometric polynomials of two variables of degree N, it suffices to take $\delta = cN^{-2}$ where c is a sufficiently small positive numerical constant. It is possible that their bounds can be significantly improved if instead of recovering the exact values of the Betti numbers one tries to recover them with a small relative error.

6. The Sketch of the Proof of the Theorem on the Number of Nodal Domains

Here, we will describe the main ideas behind the proof of Theorem 5.7. All the details can be found in [32].

6.1. The lower bound $\mathcal{E}N(f_n) \geq cn^2$. This is the simplest part of the story. Denote by $d(.,.)$ the spherical distance. Given a point $x \in \mathbb{S}^2$ and a large positive constant C, we consider the event

$$\Omega_x = \big\{ f_n(x) > C, \text{ and } f_n(y) < -C \text{ for all } y \text{ satisfying } d(x,y) = \frac{\rho}{n} \big\},$$

where ρ is a constant whose value will be specified below. Clearly, if the event Ω_x occurs, then the disk of radius ρ/n centered at x contains a closed nodal line of f_n. We claim that

$$\mathcal{P}(\Omega_x) \geq c > 0\,,$$

where c is a positive constant. The reason is rather straightforward: for every point $x \in \mathbb{S}^2$, there exists a function $b_x \in \mathcal{H}_n$ with $\|b_x\| = 1$ such that

$$b_x(x) > c_0\sqrt{n} \text{ and } b_x(y) < -c_0\sqrt{n} \text{ whenever } d(x,y) = \frac{\rho}{n}\,.$$

One can take as b_x the zonal spherical harmonic with "pole" x. Then we can represent f_n in the form

$$f_n = \xi_0 b_x + f_x$$

where ξ_0 is a Gaussian random variable with $\mathcal{E}\xi_0^2 = \frac{1}{2n+1}$, and f_x is a Gaussian spherical harmonic with $\mathcal{E}\|f_x\|^2 = \frac{2n}{2n+1}$ independent of ξ_0, and check that with positive probability, the 'perturbation' f_x cannot destroy a short nodal curve around point x provided by the function b_x.

It remains to pack the sphere $\mathbb{S}^2$ by $\simeq n^2$ disjoint disks of radius $2\rho/n$. With a positive probability, each of these disks contains a closed nodal line of f_n. Whence, the lower bound for $\mathcal{E}N(f_n)$.

6.2. Levy's concentration of measure principle. To establish the exponential concentration of the random variable $N(f_n)$ around its median, we would like to use a version of classical Levy's concentration of measure principle.

Given a set K, we denote by $K_{+\rho}$ the ρ-neighbourhood of K. We apply this notation to subsets of $\mathcal{H}_n$ and the L^2-distance, to subsets of $\mathbb{S}^2$ and the usual spherical distance, and also to subsets of $\mathbb{R}^d$ with the Euclidean distance. The following Gaussian isoperimetric theorem is due to Sudakov and Tsirelson [45] and Borell [7]:

Theorem 6.1. *Let γ_d be the standard Gaussian measure on $\mathbb{R}^d$. Let $\Sigma \subset \mathbb{R}^d$ be a Borel set, and Π be an affine half-space such that*

$$\gamma_d(\Sigma) = \gamma_d(\Pi)\,.$$

Then for each $t > 0$,

$$\gamma_d(\Sigma_{+\rho}) \geq \gamma_d(\Pi_{+\rho})\,.$$

A simple computation shows that if $\gamma_d(\Pi_{+\rho})$ is not too close to 1, then $\gamma_d(\Pi)$ must be exponentially small in d, like $\exp[-c\rho^2 d]$. Applying this to the $2n+1$-dimensional space $\mathcal{H}_n$ of spherical harmonics of degree n, we get

Corollary 6.1 (Concentration of Gaussian measure on $\mathcal{H}_n$). *Let $G \subset \mathcal{H}_n$ be any measurable set of spherical harmonics. Suppose that the set $G_{+\rho}$ satisfies $\mathcal{P}(G_{+\rho}) < \frac{3}{4}$. Then $\mathcal{P}(G) \le 2e^{-c\rho^2 n}$.*

To use the concentration of measure principle, we need to show that the number $N(f)$ doesn't change too much under slight perturbations of f in the $L^2(\mathbb{S}^2)$-norm. Certainly, this is not true for all $f \in \mathcal{H}_n$, but we will show that the "unstable" spherical harmonics $f \in \mathcal{H}_n$ for which small perturbations can lead to a drastic decrease in the number of nodal lines are exponentially rare. Here is a key lemma which is probably the most novel part of the whole story:

Lemma 6.1 (Uniform lower semi-continuity of $N(f_n)/n^2$). *For every $\varepsilon > 0$, there exist $\rho > 0$ and an exceptional set $E \subset \mathcal{H}_n$ of probability $\mathcal{P}(E) \le C(\varepsilon)e^{-c(\varepsilon)n}$ such that for all $f \in \mathcal{H}_n \setminus E$ and for all $g \in \mathcal{H}_n$ satisfying $\|g\| \le \rho$, we have*

$$N(f+g) \ge N(f) - \varepsilon n^2 .$$

The uniform lower semi-continuity lemma readily yields the exponential concentration of the random variable $N(f_n)/n^2$ near its median a_n. First, consider the set

$$G = \big\{ f \in \mathcal{H}_n \colon N(f) > (a_n + \varepsilon)n^2 \big\}.$$

Then for $f \in (G \backslash E)_{+\rho}$, we have $N(f) > a_n n^2$, and therefore, $\mathcal{P}\big((G\backslash E)_{+\rho}\big) \le \frac{1}{2}$. Hence, by the concentration of Gaussian measure, $\mathcal{P}(G \setminus E) \le 2e^{-c\rho^2 n}$, and finally,

$$\mathcal{P}(G) \le \mathcal{P}(G \setminus E) + \mathcal{P}(E) \le 2e^{-c\rho^2 n} + C(\varepsilon)e^{-c(\varepsilon)n} \le C(\varepsilon)e^{-c(\varepsilon)n} .$$

Now, we turn to the set

$$F = \big\{ f \in \mathcal{H}_n \colon N_f < (a_n - \varepsilon)n^2 \big\}.$$

Then

$$F_{+\rho} \subset \big\{ f \in \mathcal{H}_n \colon N_f < a_n n^2 \big\} \cup E ,$$

so that

$$\mathcal{P}(F_{+\rho}) \le \frac{1}{2} + C(\varepsilon)e^{-c(\varepsilon)n} < \frac{3}{4}$$

for large n, and it follows that $\mathcal{P}(F) \le 2e^{-c\rho^2 n}$.

6.3. The uniform lower continuity of the functional $f \mapsto N(f)$ outside of an exceptional set. Here we explain how we prove Lemma 6.1.

6.3.1. Exceptional spherical harmonics E with unstable nodal portraits. Instability of the nodal portrait of a spherical harmonic $f \in \mathcal{H}_n$ under small perturbations is caused by points where f and ∇f are simultaneously small. Let α and δ be small positive parameters, and let R be a large positive parameter (all of them will depend on ε from Lemma 6.1). Cover the sphere $\mathbb{S}^2$ by $\simeq R^{-2}n^2$ disks D_j of radius R/n in such a way that the concentric disks $4D_j$ with 4 times larger radius cover the sphere with a bounded multiplicity. We call the disk D_j *stable* if for each $x \in 3D_j$ either $|f(x)| \geq \alpha$ or $|\nabla f(x)| \geq \alpha n$. Otherwise, the disk D_j is *unstable*. We call the spherical harmonic $f \in \mathcal{H}_n$ *exceptional* if the number of unstable disks is at least δn^2, and denote by E the set of all exceptional spherical harmonics of degree n.

Lemma 6.2. *Given $\delta > 0$, there exist positive $C(\delta)$ and $c(\delta)$ such that*

$$\mathcal{P}(E) \leq C(\delta)e^{-c(\delta)n}$$

provided that the constant α is sufficiently small.

Curiously, the proof of this lemma uses the concentration of measure principle again. It also uses the fact that given $x \in \mathbb{S}^2$, the Gaussian random variable $f(x)$ and the Gaussian random vector $\nabla f(x)$ are independent.

6.3.2. Identification of unstable connected components. It remains to show that at most εn^2 nodal components of a stable spherical harmonic can disappear after perturbation of f by another spherical harmonic $g \in \mathcal{H}_n$ with sufficiently small L^2-norm. First, in several steps, we identify possibly 'unstable' connected components of the nodal set $Z(f)$ that can disappear after perturbation, show that their number is small compared to n^2, and discard them. Then we verify that all other connected components of $Z(f)$ do not disappear after the perturbation.

First, we discard the nodal components Γ whose diameter is bigger than R/n. By the upper bound in the length estimate in Theorem 5.3, their number is $\lesssim CR^{-1}n^2$ which is small compared to n^2.

With each remaining component Γ of the nodal set $Z(f)$ we associate a disk D_j such that $D_j \cap \Gamma \neq \emptyset$. Then $\Gamma \subset 2D_j$. Since each nodal domain contains a disk of radius c/n (Theorem 5.2), the number of components Γ intersecting D_j (and, thereby, contained in $2D_j$) is bounded.

Second, we discard the components Γ with unstable disks D_j. Since f is not exceptional, and since each disk D_j cannot intersect too many components contained in $2D_j$, the number of such components is also small compared to n^2.

At last, we discard the components Γ such that

$$\max_{3D_j} |g| \geq \alpha\,.$$

To estimate the number N of such disks, we denote by $D_j^* \subset 4D_j$ the disk of radius $1/n$ centered at the point y_j where $|g|$ attains its maximum in $3D_j$. By standard elliptic estimates,

$$\int_{D_j^*} |g|^2 \gtrsim n^{-2}|g(y_j)| = \alpha^2 n^{-2}\,,$$

whence

$$\rho^2 \geq \|g\|_{L^2(\mathbb{S}^2)} \gtrsim N\alpha^2 n^{-2}\,,$$

that is, $N \lesssim \rho^2\alpha^{-2}n^2$. As above, we conclude that the number of components Γ affected by this is $\lesssim R^2 N \lesssim R^2\rho^2\alpha^{-2}n^2$ which is much less than εn^2 provided that ρ^2 is much less than $\varepsilon\alpha^2 R^{-2}$.

6.3.3. Verification of stability of the remaining connected components. Now, we claim that the remaining components Γ cannot be affected by the perturbation of f by g. To see this, we consider the connected component $A_\Gamma(t)$ of the set $\{|f| < t\}$ that contains Γ, and look what may happen with this component when t grows from 0 to α.

- As long as $A_\Gamma(t)$ stays away from the boundary $\partial(3D_j)$, it cannot merge with another component of $\{|f| < t\}$ because such a merge can occur only at a critical point of f and there are none of them in $A_\Gamma(t) \cap 3D_j$.

- For the same reason, neither of the two boundary curves of $A_\Gamma(t)$ can collapse and disappear.

- At last, $A_\Gamma(t)$ cannot reach $\partial(3D_j)$ before it merges with some other component: indeed, if $x \in A_\Gamma(t)$ and $A_\Gamma(t)$ lies at a positive distance from the boundary $\partial(3D_j)$ then we can go from x in the direction of ∇f if $f(x) < 0$ and in the direction of $-\nabla f$ if $f(x) > 0$. In any case, since $|\nabla f| > \alpha n$ in $A_\Gamma(t)$, we shall reach the zero set $Z(f)$ after going the length $1/n$ or less. Since Γ is the only component of $Z(f)$ in $A_\Gamma(t)$ before any merges, we conclude that $A_\Gamma(t) \subset \Gamma_{+1/n}$. Recalling that $\operatorname{dist}(\Gamma, \partial(3D_j)) > R/n$, we see that, for each $t \leq \alpha$, $A_\Gamma(t)$ stays away from the boundary $\partial(3D_j)$.

Thus, each component Γ lies in a topological annulus $A_\Gamma = A_\Gamma(\alpha)$ which is contained with its boundary in the open disk $3D_j$ and such that $f = +\alpha$ in one boundary curve of A_Γ and $f = -\alpha$ on the other. Recalling that $|g| < \alpha$ in $3D_j$, we conclude that $Z(f+g)$ has at least one connected component in A_Γ.

6.4. Existence of the limit of $\mathcal{E}N(f_n)/n^2$. We already know that $\mathcal{E}N(f_n) \gtrsim n^2$ and that $N(f_n)/n^2$ concentrates near its median exponentially. Thus, to finish the proof of Theorem 5.7, it remains to show that the sequence $\{\mathcal{E}N(f_n)/n^2\}$ converges. We deduce this from the fact that the Gaussian spherical harmonic has a scaling limit combined with rotation invariance of the distribution of f_n. Since this part does not require any new ideas beyond the ones we've already introduced, we just refer the reader to [32] for the details.

6.5. Comments and questions. Making use of a non-critical version of their percolation model, Bogomolny and Schmit obtained in [9] a series of predictions for the behaviour of the components of level sets which agree with numerics. In a stark contrast, we do not have a rigorous answer even to the following most basic question:

Question 6.1. *Prove that for each $\varepsilon > 0$ and each $\eta > 0$, the probability that the level set $\{x \in \mathbb{S}^2 \colon f_n(x) > \varepsilon\}$ has a component of diameter larger than η tends to zero as $n \to \infty$.*

One of the reasons for our ignorance is the aforementioned non-locality of the number of connected components. Another essential difficulty is a very slow decay of the correlations which does not allow us to think of restrictions of our process to a collection of well-separated disks as of almost independent processes.

Question 6.2. *Estimate the mean number of large components of the nodal set whose diameter is much bigger than $1/n$. For instance, of those whose diameter is comparable to $n^{-\alpha}$ with $0 < \alpha < 1$.*

Nothing is known about the number of connected components of the nodal set for 'randomly chosen' high-energy Laplace eigenfunction f_λ on an arbitrary compact surface M without boundary endowed with a smooth Riemannian metric g. It is tempting to expect that Theorem 5.7 models what is happening when M is the two-dimensional sphere $\mathbb{S}^2$ endowed with a generic Riemannian metric g that is sufficiently close (with several derivatives) to the constant one.

Instead of perturbing the 'round metric' on the sphere $\mathbb{S}^2$, one can add a small potential V to the Laplacian on the sphere. The question remains just as hard.

Acknowledgements

We are grateful to Manjunath Krishnapur, Yuri Makarychev, Yuval Peres, Leonid Polterovich, Zeév Rudnick, Bernard Shiffman, Boris Tsirelson, Sasha Volberg, and Steve Zelditch for many helpful conversations on the subject of these notes, and to Alex Barnett, Manjunath Krishnapur, and Balint Virág for providing us with inspiring computer generated pictures.

References

[1] M. Ajtai, J. Komlós and G. Tusnády, *On optimal matchings*, Combinatorica **4** (1984), 259–264.

[2] P. Bérard, *Volume des ensembles nodaux des fonctions propres du laplacien*, Bony-Sjöstrand-Meyer seminar, 1984–1985, Exp. No. 14 , 10 pp., École Polytech., Palaiseau, 1985.

[3] M. V. Berry, *Regular and irregular semiclassical wavefunctions.* J. Phys. A **10** (1977), 2083–2091.

[4] M. V. Berry, *Statistics of nodal lines and points in chaotic quantum billiards: perimeter corrections, fluctuations, curvature*, J. Phys. A **35** (2002), 3025–3038.

[5] P. Bleher, B. Shiffman, S. Zelditch, *Universality and scaling of correlations between zeros on complex manifolds*, Invent. Math. **142** (2000), 351–395. `arXiv:math-ph/9904020`

[6] G. Blum, S. Gnutzmann, U. Smilansky, *Nodal Domains Statistics: A Criterion for Quantum Chaos*. Phys. Rev. Letters, **88** (2002), 114101. `arXiv:nlin/0109029v1`

[7] Chr. Borell, *The Brunn-Minkowski inequality in Gauss space*. Invent. Math. **30** (1975), 207–216.

[8] E. Bogomolny, C. Schmit, *Percolation Model for Nodal Domains of Chaotic Wave Functions*. Phys. Rev. Letters, **88** (2002), 114102. `arXiv:nlin/0110019v1`

[9] E. Bogomolny, C. Schmit, *Random wavefunctions and percolation*, J. Phys. A **40** (2007), 14033–14043. `arXiv:0708.4335v1`

[10] J. Brüning, *Über Knoten von Eigenfunktionen des Laplace-Beltrami-Operators*, Math. Z. **158** (1978), 15–21.

[11] S. Chatterjee, R. Peled, Y. Peres, D. Romik, *Gravitational allocation to Poisson points*, Annals of Math., to appear. `arXiv:math/0611886`; *Phase Transitions in Gravitational Allocation*, Geom. and Funct. Anal., to appear. `arXiv:0903.4647`

[12] H. Donnelly, Ch. Fefferman, *Nodal sets of eigenfunctions on Riemannian manifolds*, Invent. Math. **93** (1988), 161–183.

[13] H. Donnelly, Ch. Fefferman, *Nodal sets for eigenfunctions of the Laplacian on surfaces*, J. Amer. Math. Soc. bf 3 (1990), 333–353.

[14] H. Donnelly, Ch. Fefferman, *Growth and geometry of eigenfunctions of the Laplacian*, Analysis and partial differential equations, 635–655, Lecture Notes in Pure and Appl. Math., 122, Dekker, New York, 1990.

[15] M. Douglas, B. Shiffman, and S. Zelditch, *Critical points and supersymmetric vacua. I*, Comm. Math. Phys. **252** (2004), 325–358. `arxiv: math.CV/0402326`

[16] A. Eremenko, D. Jakobson, N. Nadirashvili, *On nodal sets and nodal domains on* $\mathbb{S}^2$ *and* $\mathbb{R}^2$, Ann. Inst. Fourier (Grenoble) **57** (2007), 2345–2360. `arXiv:math/0611627v2`

[17] P. J. Forrester and G. Honner, *Exact statistical properties of the zeros of complex random polynomials.* J. Phys. A **32** (1999), 2961–2981. `arXiv:cond-mat/9811142`

[18] B. Jancovici, J. L. Lebowitz, G. Manificat, *Large charge fluctuations in classical Coulomb systems*. J. Statist. Phys. **72** (1993), 773–787.

[19] A. E. Holroyd, R. Pemantle, Y. Peres, O. Schramm, *Poisson Matching*, Ann. Inst. Henri Poincaré Probab. Stat. **45** (2009), 266–287. arXiv:0712.1867

[20] C. Hoffman, A. E. Holroyd and Y. Peres, *A Stable Marriage of Poisson and Lebesgue*, A stable marriage of Poisson and Lebesgue. Ann. Probab. **34** (2006), 1241–1272. arxiv:math.PR/0505668; *Tail Bounds for the Stable Marriage of Poisson and Lebesgue*, arxiv: math.PR/0507324.

[21] B. Hough, M. Krishnapur, Y. Peres, B. Virág, *Zeros of Gaussian Analytic Functions and Determinantal Point Processes, Amer. Math. Soc.*, 2009. Electronic version available at stat-www.berkeley.edu/~peres/GAF_book.pdf

[22] M. Krishnapur, *Overcrowding estimates for zeroes of Planar and Hyperbolic Gaussian analytic functions*, J. Statist. Phys. **124** (2006), 1399–1423. arxiv: math.PR/0510588

[23] M. Laczkovich, *Equidecomposability and discrepancy; a solution of Tarski's circle-squaring problem*, J. Reine Angew. Math. **404** (1990).

[24] M. Laczkovich, *Uniformly spread discrete sets in* $\mathbb{R}^d$, J. London Math. Soc. (2) **46** (1992).

[25] T. Leighton and P. Shor, *Tight bounds for minimax grid matching with applications to the average case analysis of algorithms*, Combinatorica **9** (1989), 161–187.

[26] H. Lewy, *On the minimum number of domains in which the nodal lines of spherical harmonics divide the sphere.* Comm. Partial Differential Equations **2** (1977), 1233–1244.

[27] M. S. Longuet-Higgins, *The statistical analysis of a random, moving surface.* Philos. Trans. Roy. Soc. London Ser. A. **249** (1957), 321–387. *Statistical properties of an isotropic random surface.* Philos. Trans. Roy. Soc. London. Ser. A. **250** (1957), 157–174. *The statistical geometry of random surfaces.* 1962 Proc. Sympos. Appl. Math., Vol. XIII pp. 105–143 American Mathematical Society, Providence, R.I.

[28] T. L. Malevich, *Contours that arise when the zero level is crossed by Gaussian fields.* Izv. Acad. Nauk Uzbek. SSR **16** (1972), no 5, 20–23. (Russian)

[29] D. Mangoubi, *On the inner radius of a nodal domain*, Canad. Math. Bull. **51** (2008), 249–260. arXiv:math/0511329v3 *The Volume of a Local Nodal Domain*, arXiv:0806.3327v4

[30] K. Mischaikov, T. Wanner, *Probabilistic validation of homology computations for nodal domains* Ann. Appl. Probab. **17** (2007), 980–1018.

[31] F. Nazarov, L. Polterovich, M. Sodin, *Sign and area in nodal geometry of Laplace eigenfunctions*, Amer. J. Math. **127** (2005), 879–910. arXiv:math/0402412v2

[32] F. Nazarov, M. Sodin, *On the number of nodal domains of random spherical harmonics.* Amer. J. Math. **131** (2009), 1337–1357. arXiv:0706.2409v1

[33] F. Nazarov and M. Sodin, *What is a . . . Gaussian entire function*? Notices Amer. Math. Soc., March, 2010.

[34] F. Nazarov and M. Sodin, *Fluctuations in random complex zeroes*, arXiv:1003.4251v1.

[35] F. Nazarov and M. Sodin, *Correlation functions for random complex zeroes: strong clustering and weak universality*, In preparation.

[36] F. Nazarov, M. Sodin, and A. Volberg, *Transportation to random zeroes by the gradient flow*, Geom. and Funct. Anal. **17** (2007), 887–935. arXiv:math/0510654

[37] F. Nazarov, M. Sodin, and A. Volberg, *The Jancovici - Lebowitz - Manificat law for large fluctuations of random complex zeroes*, Commun. Math. Phys. **284** (2008), 833–865. arXiv:0707.3863

[38] A. Nishry, *Asymptotics of the Hole Probability for Zeros of Random Entire Functions*, Int. Math. Res. Not. IMRN, to appear. arXiv:0903.4970

[39] A. Nishry, *The Hole Probability for Gaussian Entire Functions*, Israel J. Math., to appear. arXiv:0909.1270

[40] F. Oravecz, Z. Rudnick, I. Wigman, *The Leray measure of nodal sets for random eigenfunctions on the torus*, Ann. Inst. Fourier (Grenoble) **58** (2008), 299–335. arXiv:math-ph/0609072v2

[41] A. Pleijel, *Remarks on Courant's nodal line theorem*, Comm. Pure Appl. Math. **9** (1956), 543–550.

[42] B. Rider and B. Virág, *The noise in the circular law and the Gaussian free field*, Int. Math. Res. Not. IMRN, no 2 (2007), arXiv:math/0606663; *Complex determinantal processes and H^1 noise*, Electron. J. Probab. **12** (2007), 1238–1257, arXiv:math/0608785

[43] Z. Rudnick, I. Wigman, *On the volume of nodal sets for eigenfunctions of the Laplacian on the torus*, Ann. Henri Poincaré **9** (2008), 109–130. arXiv:math-ph/0702081v2

[44] M. Sodin, *Zeroes of Gaussian analytic functions*, European Congress of Mathematics, 445–458, Eur. Math. Soc., Zürich, 2005. arXiv:math/0410343

[45] M. Sodin and B. Tsirelson, *Random complex zeroes.* I. *Asympotic normality*, Israel J. Math. **144** (2004), 125–149; II. *Perturbed Lattice*, ibid **152** (2006), 105–124; III. *Decay of the hole probability*, ibid **147** (2005), 371–379. arxiv:math.CV/0210090, math.CV/0309449, and math.CV/0312258

[46] M. Sodin and B. Tsirelson, *Uniformly spread measures and vector fields*, arXiv:0801.2505

[47] A. Soshnikov, *Gaussian Limit for Determinantal Random Point Fields*, Annals of Probab. **30** (2002), 171–187. arXiv:math/0006037

[48] V. N. Sudakov, B. S. Cirel'son, *Extremal properties of half-spaces for spherically invariant measures.* Problems in the theory of probability distributions, II. Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **41** (1974), 14–24. (Russian)

[49] P. Swerling, *Statistical properties of the countours of random surfaces.* IRE Trans. Inf. Theory, **8** (1962), 315–321.

[50] G. Szegö, *Orthogonal polynomials.* Fourth edition. American Mathematical Society, Colloquium Publications, Vol. XXIII. American Mathematical Society, Providence, R.I., 1975.

[51] M. Talagrand, *Matching theorems and empirical discrepancy computations using majorizing measures*, J. Amer. Math. Soc. **7** (1994), 455–537.

[52] B. Tsirelson, *Moderate deviations for random fields and random complex zeroes*, arXiv:0801.1050v1

[53] I. Wigman, *Fluctuations of the nodal length of random spherical harmonics*, arXiv:0907.1648v1

[54] S. Zelditch, *From random polynomials to symplectic geometry*, XIIIth International Congress on Mathematical Physics (London, 2000), 367–376, Int. Press, Boston, MA, 2001; *Asymptotics of polynomials and eigenfunctions*, Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002), 733–742, Higher Ed. Press, Beijing, 2002.

[55] S. Zelditch, *Real and complex zeros of Riemannian random waves*, Spectral analysis in geometry and number theory, 321–342, Contemp. Math., 484, Amer. Math. Soc., Providence, RI, 2009. arXiv:0803.4334

[56] S. Zelditch, *Local and global analysis of eigenfunctions*, Advanced Lectures in Mathematics (ALM) **7**, 545–658 (2008) arXiv:0903.3420v1

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Potential Analysis Meets Geometric Measure Theory

Tatiana Toro*

Abstract

A central question in Potential Theory is the extent to which the geometry of a domain influences the boundary regularity of solutions to divergence form elliptic operators. To answer this question one studies the properties of the corresponding elliptic measure. On the other hand one of the central questions in Geometric Measure Theory (GMT) is the extent to which the *regularity* of a measure determines the geometry of its support. The goal of this paper is to present a few instances in which techniques from GMT and Harmonic Analysis come together to produce new results in both of these areas. In particular, the work described in section 3 makes it clear that for this type of problems in higher dimensions, GMT is the right alternative to complex analysis in dimension 2.

Mathematics Subject Classification (2010). Primary 28A33; Secondary 31A15.

Keywords. Elliptic measure, Harmonic measure, Ahlfors regular.

1. Introduction

A central theme in potential theory is understanding the properties of solutions to divergence form elliptic operators. On regular domains one can associate to such operators a family of probability measures indexed by the points in the domain. All the measures in this family are mutually continuous (in fact they are A_∞ weights with respect to each other). We refer to any one of them as the elliptic measure. To address the question of boundary regularity of the

*The author was partially supported by NSF grants DMS-0600915 and DMS-0856687
Department of Mathematics, University of Washington, Seattle, WA 98195-4350.
E-mail: toro@math.washington.edu.

solutions to these operators one studies the properties of the corresponding elliptic measure. The introduction of techniques from geometric measure theory has enabled us to deepen our understanding of the subject in two different directions. On the one hand we have begun to understand the properties of the elliptic measure on rough domains. On the other hand we have studied to extent to which the *regularity* of the harmonic measure (that is the elliptic measure associated to the Laplacian) determines the geometry of the boundary of a domain. The underlying trend is that there is a strong relationship between the *regularity* properties of the elliptic measure and the geometry of the domain. Roughly, the geometry of the domain determines the doubling properties of the elliptic measure as well as its behavior with respect to the surface measure of the boundary in the cases where this notion makes sense, and vice versa. While the two problems above are very different in nature, they share common features. The key to this work is the interplay between harmonic analysis and geometric measure theory.

An important initial contribution of GMT to the calculus of variations and geometric analysis was the introduction of large classes of geometric objects (e.g. sets of locally finite perimeter, rectifiable sets (see [EG])) regular enough to be admissible candidates in minimizations problems but without any, a priori, classical smoothness properties. Quantified versions of these notions (see [DS]) allow us to introduce new classes of domains on which the boundary regularity of solutions to divergence form elliptic operators can be well understood. With the appropriate notion of convergence, these domains appear naturally as limits of smooth domains. The first results in this direction concerned the doubling properties of the harmonic measure and the regularity of the Poisson kernel (i.e. the Radon-Nikodym derivative of harmonic measure with respect to surface measure to the boundary) on rough domains, see [JK], [S1], [DJ], [KT1], [HMT]. The results presented in section 2 include recent work on the regularity properties of the elliptic measure corresponding to operators which are perturbations of the Laplacian on rough domains. The crucial idea behind this work is that when looked at with the right lens from harmonic analysis, chord arc domains (see the definition in section 2) have many features in common with the half space.

The converse problem addresses the question of whether the properties of the harmonic measure determine the geometry of the domain. In the plane this problem as been studied extensively (see section 3 for a brief summary). In higher dimensions this problem was initially studied as a free boundary regularity problem for the harmonic measure and the Poisson kernel (see [KT2], [KT3]). Recently combining a monotonicity inequality from free boundary regularity problems ([ACF]) and the machinery of tangent measures from geometric measure theory ([P]) we give a full description of the boundary of a domain in terms of the harmonic measure (see section 3, [KPT]). GMT accomplishes in higher dimensions what complex analysis did in the plane.

2. Regularity of Elliptic Measure on Rough Domains

The Dirichlet problem addresses the following question: given a bounded domain $\Omega \subset \mathbb{R}^n$ and a function $f \in C(\partial\Omega)$ does there exists u satisfying

$$\begin{cases} Lu = \operatorname{div}(A(X)\nabla u) &= 0 \text{ in } \Omega \\ u &= f \text{ on } \partial\Omega \ ? \end{cases} \tag{1}$$

Here $A(X) = (a_{ij}(X))$ is a symmetric measurable matrix such that for all $X \in \Omega$ and $\xi \in \mathbb{R}^n$, $\lambda|\xi|^2 \leq \sum_{i,j=1}^n a_{ij}(X)\xi_i\xi_j \leq \Lambda|\xi|^2$, i.e. L is strongly elliptic.

The domain Ω is regular for L, if for all $f \in C(\partial\Omega)$, there exists $u_f = u \in C(\overline{\Omega})$ satisfying (1). Moreover Ω is regular for L if and only if it is regular for the Laplacian Δ (i.e. when A is the identity matrix). If Ω is regular the maximum principle and the Riesz representation theorem ensure that there is a family of probability measures $\{\omega_L^X\}_{X\in\Omega}$ such that

$$u(X) = \int_{\partial\Omega} f(Q)d\omega_L^X(Q),$$

ω_L^X is called the $L-$elliptic measure of Ω with pole X. Jerison and Kenig introduced a special class of regular domains, the non-tangentially accessible (NTA) domains (see [JK]). They proved that on NTA domains the elliptic measure is doubling and that the non-tangential limit of the solution of (1) at the boundary exists and coincides with f ω_L a.e.. Quasispheres (i.e. images of the unit ball by quasi-conformal transformations of Euclidean space) and Lipschitz domains are both examples of NTA domains. In the case that Ω is a Lipschitz domain (i.e. it can be locally represented a the region above the graph of a Lipschitz function) a natural question arises: what is the relationship between ω_L and the surface measure of $\partial\Omega$, $\sigma = \mathcal{H}^{n-1} \llcorner \partial\Omega$? Here $\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure. This question also makes sense in a larger class of domains, that of chord arc domains (see also SKT domains in [HMT]). A chord arc domain (CAD) is an NTA domain whose surface measure at the boundary is Ahlfors regular, i.e. there exists $C > 1$ such that for all $Q \in \partial\Omega$ and $r \in (0, \operatorname{diam}\Omega)$

$$C^{-1}r^{n-1} \leq \sigma(B(Q,r)) \leq Cr^{n-1}. \tag{2}$$

The best such C is called the Ahlfors regularity constant of σ.

We address the following question. Let Ω be a CAD, L an operator as above and u the solution of (1). Given $1 < p < \infty$ is the non-tangential maximal function of u bounded in $L^p(\sigma)$? Namely we are interested in the $(D)_p$ problem, i.e. given $f \in C(\partial\Omega)$ and u satisfying

$$\begin{cases} Lu &= 0 \text{ in } \Omega \\ u &= f \text{ on } \partial\Omega \end{cases}$$

does there exist $C > 0$ depending only on the geometry of the domain, the Ahlfors regularity constant of σ and the ellipticity constants of L such that following bound hold?

$$\|N(u)\|_{L^p(\sigma)} \leq C\|f\|_{L^p(\sigma)}, \tag{3}$$

where the non-tangential maximal function of u is defined for $Q \in \partial\Omega$ by

$$N(u)(Q) = \sup_{X\in\Gamma(Q)} |u(X)|$$

$$\Gamma(Q) = \{X \in \Omega : |X - Q| < 2\delta(X)\} \text{ and } \delta(X) = \operatorname{dist}(\mathrm{X}, \partial\Omega). \tag{4}$$

This PDE question has an equivalent formulation in harmonic analysis, by means of the theory of weights. In fact (3) holds if and only if the elliptic measure of L, ω_L and σ are mutually absolutely continuous and the Radon-Nikodym derivative $k_L = \frac{d\omega_L}{d\sigma}$ satisfies the reverse Hölder inequality:

$$\left(\frac{1}{\sigma(\Delta(Q,r))}\int_{\Delta(Q,r)} k_L^q\, d\sigma\right)^{\frac{1}{q}} \leq C\frac{1}{\sigma(\Delta(Q,r))}\int_{\Delta(Q,r)} k_L\, d\sigma, \tag{5}$$

where $\frac{1}{p} + \frac{1}{q} = 1$, $\Delta(Q,r) = B(Q,r) \cap \partial\Omega$, $Q \in \partial\Omega$ and $r \in (0, \operatorname{diam}\Omega)$. If (5) holds we say that $\omega_L \in B_q(\sigma)$.

We briefly recall the history of the $(D)_p$ problem. Dahlberg's pioneering work established that for the Laplacian, i.e. $L = \Delta$, the harmonic measure $\omega_\Delta = \omega$ satisfies $\omega \in B_2(\sigma)$, thus $(D)_2$ holds for all Lipschitz domains [D1]. By contrast if Ω is a CAD there exists $q \in (1,\infty)$ so that $\omega \in B_q(\sigma)$, but such q is not uniform across the class of chord arc domains (see [DJ] and [S1]). It was shown in [CFK], and independently in [MM], that there are operators L for which ω_L and σ are mutually singular, hence neither (3) nor (5) hold. Thus one of the main questions in this area is to find sharp conditions that ensure that (3) and (5) are satisfied.

Suppose that we have two operators L_0 and L_1, whose respective coefficient matrices A_0 and A_1 coincide on a neighborhood of $\partial\Omega$. Then if $(D)_p$ holds for L_0 it also holds for L_1. This is a consequence of the properties of non-negative harmonic functions on NTA domains. Thus $(D)_p$ is a property that only depends on the behavior of the coefficients of L near $\partial\Omega$. Therefore we are lead to consider the following notion: we say that L_1 is a perturbation of L_0 if there exists a constant $C > 0$ such that the deviation function

$$a(X) = \sup\{|A_1(Y) - A_0(Y)| : Y \in B(X, \delta(X)/2)\} \tag{6}$$

satisfies

$$\sup_{0<r<\operatorname{diam}\Omega,\, Q\in\partial\Omega} h(Q,r) \leq C, \tag{7}$$

where

$$h(Q,r) = \left(\frac{1}{\sigma(\Delta(Q,r))}\int_{T(\Delta(Q,r))} \frac{a^2(X)}{\delta(X)}\, dX\right)^{1/2}, \tag{8}$$

i.e. $\frac{a^2(X)}{\delta(X)}\,dX$ is a Carleson measure. Here $T(\Delta(Q,r)) = B(Q,r)\cap\Omega$ is the Carleson region associated to the surface ball $\Delta(Q,r)$. Note that if (7) holds then $L_1 = L_0$ on $\partial\Omega$. We include below some of the most remarkable results in this direction. In the sequel we assume that $0\in\Omega$ and we denote by ω_i the L_i-elliptic measure in Ω with pole 0.

Theorem 2.1. *[D2] Let $\Omega = B(0,1)$. If $L_0 = \Delta$, and $\lim_{r\to 0}\sup_{|Q|=1} h(Q,r) = 0$, then $\omega_1 \in B_q(\sigma)$ for all $q > 1$.*

In [F], Fefferman removed the smallness condition of $h(Q,r)$ above by defining an appropriate quantity $A(a)(Q)$. Recall that $A_\infty(\sigma) = \cup_{q>1} B_q(\sigma)$.

Theorem 2.2. *[F] Let $\Omega = B(0,1)$, $L_0 = \Delta$, $\Gamma(Q)$ be as in (4), and $A(a)(Q) = (\int_{\Gamma(Q)} \frac{a^2(X)}{\delta(X)^n} dX)^{1/2}$. If $A(a)\in L^\infty(\sigma)$ then $\omega_1 \in A_\infty(\sigma)$.*

Theorem 2.3. *[FKP] Let Ω be a Lipschitz domain. Let L_1 be such that (7) holds then $\omega_1 \in A_\infty(d\sigma)$ whenever $\omega_0 \in A_\infty(d\sigma)$.*

Theorem 2.4. *[FKP] Let Ω be a Lipschitz domain. Let G_0 denote the Green function for L_0 in Ω with pole at $0\in\Omega$. There exists an $\epsilon_0 > 0$, such that if*

$$\sup_{\Delta\subseteq\partial\Omega}\left(\frac{1}{\omega_0(\Delta)}\int_{T(\Delta)} a^2(X)\frac{G_0(X)}{\delta^2(X)}dX\right)^{1/2} \le \epsilon_0, \tag{9}$$

then $\omega_1 \in B_2(\omega_0)$.

Theorems 2.1 and 2.2 are proved using Dahlberg's idea of introducing a differential inequality to estimate the B_q norm for a family of elliptic measures. Theorem 2.3 was proved by a direct method which used Theorem 2.4. Theorem 2.4 relied on techniques from harmonic analysis.The basic approach was to look at the solution of (1) for L_1 as a perturbation of the solution of (1) for L_0 and estimate the *error* term using the duality properties of the tent spaces introduced in [CMS].Tent spaces were initially defined as special subspaces of functions on the half space. The geometry of a Lipschitz domain and the properties of the surface measure of its boundary allow the notion of tent space to be extended to this class of domains. Let $\Omega\subset\mathbb{R}^n$ be a Lipschitz domain, for $1\le p\le\infty$ the tent space $\mathcal{T}^p$ is defined by

$$\mathcal{T}^p = \left\{f\in L^2(\Omega):\ A(f)\in L^p(\sigma)\right\} \tag{10}$$

where

$$A(f)(Q) = \left(\int_{\Gamma(Q)}\frac{f^2(X)}{\delta(X)^n}dX\right)^{1/2}, \tag{11}$$

and

$$\mathcal{T}^\infty = \left\{f\in L^2(\Omega):\ C(f)\in L^\infty(\sigma)\right\} \tag{12}$$

where

$$C(f)(Q) = \sup_{Q\in\Delta}\left(\frac{1}{\sigma(\Delta)}\int_{T(\Delta)}\frac{f^2(X)}{\delta(X)}\,dX\right)^{1/2}. \tag{13}$$

As mentioned above the proof of Theorem 2.4 relies on the duality of tent spaces which is expressed as $\mathcal{T}^q$ is the dual of $\mathcal{T}^p$ where $\frac{1}{p}+\frac{1}{q}=1$ for $1\leq p\leq\infty$; the $L^p(\sigma)$ equivalence of $C(f)$ and $A(f)$ for $2<p<\infty$, i.e. $\|C(f)\|_{L^p(\sigma)}\sim\|A(f)\|_{L^p(\sigma)}$; and the properties of non-negative solutions to (1) on Lipschitz domains. A careful look at the proof shows that, from the PDE point of view, only the properties of non-negative solutions to (1) on NTA domains are used. The proof of the two other properties appears to depend heavily on the geometry of the domain, in particular, on the fact that locally, truncated cones of a given direction and a given aperture with vertex at the boundary lie inside the domain. On the other hand the definitions in (10) and (12) as well as condition (9) make sense when Ω is a CAD. In recent work with Milakis and Pipher we prove that both the duality statement and the equivalence $C(f)$ and $A(f)$ in $L^p(\sigma)$ for $2<p<\infty$ hold on chord arc domains. Furthermore we prove that theorems 2.3 and 2.4 also hold on chord arc domains. A thorough understanding of the geometry of these domains is required to accomplish this.

Theorem 2.5. *[MPT] Let Ω be a CAD, there exists $\epsilon_0>0$ such that if (9) holds then $\omega_1\in B_2(\omega_0)$.*

Corollary 2.1. *[MPT] Let Ω be a CAD, there exists $\epsilon>0$ such that if $\sup_{Q\in\partial\Omega,\ r>0}h(Q,r)<\epsilon$ then $\omega_1\in A_\infty(\sigma)$ whenever $\omega_0\in A_\infty(\sigma)$.*

The following three questions, which were motivated by the results above, are currently under investigation.

- *Let Ω be a CAD. Suppose that $\sup_{Q\in\partial\Omega}\sup_{r>0}h(Q,r)<\infty$. Does $\omega_0\in A_\infty(\sigma)$ imply that $\omega_1\in A_\infty(\sigma)$?*
- *Let Ω be a CAD. Suppose that $\lim_{r\to0}\sup_{Q\in\partial\Omega}h(Q,r)=0$. If $\log k_0\in VMO(\sigma)$ does $\log k_1\in VMO(\sigma)$ where $k_j=\frac{d\omega_j}{d\sigma}$?*

  This question is motivated by the corresponding result on Lipschitz domains proved in [E] using the Dalhberg's differential inequality idea. In particular it is known that if Ω is a CAD whose unit normal is in $VMO(\sigma)$ then $\log k_0\in VMO(\sigma)$ for $L_0=\Delta$ (see [KT1]).
- *Let Ω be a CAD. Is the solvability of an endpoint BMO Dirichlet problem for a strongly elliptic operator L equivalent to $\omega_L\in A_\infty(\sigma)$?*

  This question is motivated by recent work in [DKP], where the corresponding result for Lipschitz domains was established.

The proof of theorems 2.1 and 2.2 rely to some extent on the ability to approximate Lipschitz domains by smooth interior domains in such a way that

the surface measure and the unit normal to the boundary of the original domain are the limits of the surface measures and the unit normals of the approximating domains. Such an approximation is not known to exist for chord arc domains, which raises an interesting question.

Question 2.1. *Let Ω be a CAD. Does there exist a family of smooth domains $\{\Omega_m\}_m$ such that $\Omega_m \subset \Omega$ is CAD with constants that only depend on the NTA and Ahlfors regularity constants of Ω, and $\chi_{\Omega_m} \to \chi_\Omega$ in BV_{loc}?*

We turn our attention to the Neumann and regularity problems on CAD, which is an area that has not been explored yet (see [KP] for the corresponding results on Lipschitz domains).

We say that the regularity problem for L with data in $W^{1,p}(\sigma)$ is solvable (i.e. $(R)_p$ holds) if given $f \in C(\partial\Omega) \cap W^{1,p}(\sigma)$ the corresponding u satisfying

$$\left\{ \begin{array}{rcl} Lu & = & 0 \text{ in } \Omega \\ u & = & f \text{ on } \partial\Omega \end{array} \right.$$

verifies

$$\|\widetilde{N}(\nabla u)\|_{L^p(\sigma)} \leq C\|f\|_{W^{1,p}(\sigma)}, \tag{14}$$

where $C > 0$ depends only on the geometry of the domain, the Ahlfors regularity constant of σ and the ellipticity constants of L. $\widetilde{N}$ is a modified non-tangential maximal function, introduced to overcome the fact that $\nabla u \notin L^\infty_{loc}$ in general. It is defined by

$$\widetilde{N}(F)(Q) = \sup_{X \in \Gamma(Q)} \left(\fint_{B(X,\delta(X)/2)} F^2(Z) dZ \right)^{1/2}. \tag{15}$$

We say that the Neumann problem for L with data in $L^p(\sigma)$ is solvable (i.e. $(N)_p$ holds) if given $f \in L^2(\sigma) \cap L^p(\sigma)$ with $\int_{\partial\Omega} f \, d\sigma = 0$ the corresponding u satisfying

$$\left\{ \begin{array}{rcl} Lu & = & 0 \text{ in } \Omega \\ A\nabla u \cdot \overrightarrow{n} & = & f \text{ on } \partial\Omega \end{array} \right.$$

verifies

$$\|\widetilde{N}(\nabla u)\|_{L^p(\sigma)} \leq C\|f\|_{L^p(\sigma)}, \tag{16}$$

where $C > 0$ depends only on the geometry of the domain, the Ahlfors regularity constant of σ and the ellipticity constants of L. Here $\overrightarrow{n}(Q)$ denotes the inward unit normal to $\partial\Omega$.

Question 2.2. *Let Ω be a CAD. Does $(R)_p$ hold for the Laplacian for some $p > 1$? Does $(N)_p$ hold for the Laplacian for some $p > 1$?*

In [HMT] it was proved that, for the Laplacian, given $p > 1$ there exits $\epsilon > 0$ such that if Ω is a CAD and $\overrightarrow{n} \in BMO(\sigma)$ with $\|\overrightarrow{n}\|_{BMO(\sigma)} < \epsilon$ then $(R)_p$ and $(N)_p$ hold. Locally the boundary of this type of domains can be

represented as the graph of a Lipschitz function with small constant union a set of very small measure (see [S2]). The proof in [HMT] depends heavily on this structure which is not shared by general CAD.

Two central questions in this area are:

Question 2.3. *Let Ω be a CAD. Suppose that $\sup_{Q\in\partial\Omega}\sup_{r>0} h(Q,r) < \infty$. If $(R)_{q_0}$ holds for L_0 for some $q_0 > 0$, does $(R)_{q_1}$ also hold for L_1 for some $q_1 > 0$?*

Question 2.4. *Let Ω be a CAD. Suppose that $\lim_{r\to 0}\sup_{Q\in\partial\Omega} h(Q,r) = 0$. If $(R)_q$ and $(N)_q$ hold for L_0 for some $q > 0$ do $(R)_q$ and $(N)_q$ also hold for L_1?*

3. Boundary Structure and Size Are Determined by Harmonic Measure

As mentioned in the introduction this problem is to some extent the converse of the one discussed above in the sense that the properties of the harmonic measure determine the size of the boundary and the geometry of the domain. Let us briefly describe some of the classical 2-dimensional results concerning the Hausdorff dimension of the harmonic measure and the structure of the boundary as determined by the harmonic measure. Recall that the Hausdorff dimension of ω (denote by $\mathcal{H}-\dim\omega$) is defined by

$$\begin{aligned}\mathcal{H}-\dim\omega \;=\; &\inf\{k:\ \text{there exists } E\subset\partial\Omega \text{ with } \mathcal{H}^k(E)=0 \text{ and}\\ &\omega(E\cap K)=\omega(\partial\Omega\cap K) \text{ for all compact sets } K\subset\mathbb{R}^n\}\end{aligned}\tag{17}$$

Let $\Omega\subset\mathbb{R}^2$ be a regular domain, and let ω be the harmonic measure of Ω. Carleson ([C]) and Jones & Wolff ([JW]) showed that $\mathcal{H}-\dim\omega\le 1$. If Ω is simply connected, Makarov ([Mak]) proved Oskendal's conjecture in dimension 2, i.e. $\mathcal{H}-\dim\omega = 1$.

Let $\Omega\subset\mathbb{R}^2$ be a Jordan domain (i.e. a simply connected domain bounded by a Jordan curve). A combination of the works of Choi, Makarov, McMillan and Pommerenke (see [GM] for references) shows that the boundary of Ω can be decomposed as the union of a "good set" where ω and $\mathcal{H}^1$ are mutually absolutely continuous, a set of harmonic measure 0 and a set of 1-Hausdorff measure 0, i.e.

$$\partial\Omega = G\cup S\cup N \ \text{ where } \ \omega\ll\mathcal{H}^1\ll\omega \ \text{ in } \ G,\ \omega(N)=0 \ \text{ and } \ \mathcal{H}^1(S)=0.\tag{18}$$

T. Wolff [W] showed, by a deep example, that, for $n\ge 3$, Oksendal's conjecture ($\mathcal{H}-\dim\omega = n-1$) fails. He constructed what we will call "Wolff snowflakes", domains in $\mathbb{R}^3$; for which $\mathcal{H}-\dim\omega > 2$ and others for which $\mathcal{H}-\dim\omega < 2$. In Wolff's construction, the domains are 2-sided NTA (i.e. Ω and $\text{int}(\Omega^c)$ are both NTA) which should be understood as a weak regularity

property. This plays an important role in his estimates. Whenever we refer to a "Wolff snowflake," we will mean a 2-sided NTA domain in $\mathbb{R}^n$, for which $\mathcal{H} - \dim \omega \neq n-1$. In [LVV], Lewis, Verchota and Vogel reexamined Wolff's construction and were able to produce "Wolff snowflakes" in $\mathbb{R}^n$, $n \geq 3$, for which

$$\mathcal{H} - \dim \omega^{\pm} > n-1 \quad \text{or} \quad \mathcal{H} - \dim \omega^{\pm} < n-1. \tag{19}$$

Here $\omega^{\pm}$ denotes the harmonic measure of $\Omega^{\pm}$, where $\Omega^{+} = \Omega$ and $\Omega^{-} = \text{int}(\Omega^c)$. They also observed, as a consequence of the monotonicity formula in [ACF], that if

$$\omega^{+} \ll \omega^{-} \ll \omega^{+} \quad \text{then} \quad \mathcal{H} - \dim \omega^{\pm} \geq n-1. \tag{20}$$

Returning to the case of $n = 2$, when Ω is a Jordan domain, the work of Bishop, Carleson, Garnett & Jones [BCGJ] combined with (18) yields the following new decomposition in terms of ω^{+} and ω^{-}

$$\partial\Omega = G \cup S \cup N \quad \text{where} \quad \omega^{+} \ll \omega^{-} \ll \omega^{+} \quad \text{in} \quad G,\ \omega^{\pm}(N) = 0,\ \omega^{+} \perp \omega^{-} \quad \text{on} \quad S \tag{21}$$

In [B], motivated by this last result, Bishop asked whether in the case of $\mathbb{R}^n$, $n \geq 3$, if ω^{-}, ω^{+} are mutually absolutely continuous on a set $E \subset \partial\Omega$, $\omega^{\pm}(E) > 0$, then $\omega^{\pm}$ are also mutually absolutely continuous with respect to $\mathcal{H}^{n-1}$ on E (modulo a set of $\omega^{\pm}$ measure zero) and hence $\dim_{\mathcal{H}}(E) = n-1$. Here $\dim_{\mathcal{H}}$ denotes the Hausdorff dimension. On the other hand, Lewis, Verchota and Vogel [LVV] conjectured that there are "Wolff snowflakes" in $\mathbb{R}^n$, $n \geq 3$ with $\mathcal{H} - \dim \omega^{\pm} > n-1$, for which ω^{+}, ω^{-} are not mutually singular.

In [KPT] we study these and related questions for domains $\Omega \subset \mathbb{R}^n$ which verify the weak regularity assumption of being 2-sided locally NTA (a condition satisfied by the Wolff snowflakes constructed both by Wolff and Lewis, Verchota & Vogel).

Theorem 3.1. *[KPT] Let $\Omega \subset \mathbb{R}^n$ be a 2-sided locally NTA domain. Then the boundary of Ω can be decomposed as follows:*

$$\partial\Omega = G \cup S \cup N \quad \textit{where} \quad \omega^{+} \ll \omega^{-} \ll \omega^{+} \quad \textit{in} \quad G,\ \omega^{\pm}(N) = 0,\ \omega^{+} \perp \omega^{-} \quad \textit{on} \quad S. \tag{22}$$

Moreover $\dim_{\mathcal{H}} G \leq n-1$, and if $\omega^{\pm}(G) > 0$ then $\dim_{\mathcal{H}} G = n-1$. Furthermore if $\mathcal{H}^{n-1} \llcorner \partial\Omega$ is a Radon measure then G is $(n-1)$-rectifiable.

The following theorem, which is a corollary of Theorem 3.1 proves that there are no Wolff snowflakes for which ω^{+} and ω^{-} are mutually absolutely continuous, answering a question in [LVV].

Theorem 3.2. *[KPT] Let Ω be a 2-sided locally NTA domain. Assume that ω^{+} and ω^{-} are mutually absolutely continuous, then*

$$\mathcal{H} - \dim \omega^{+} = \mathcal{H} - \dim \omega^{-} = n-1. \tag{23}$$

The study of these questions requires three main ingredients:

1. **Alt-Caffarelli-Friedman monotonicity formula:** (see [ACF]) this yields Beurling's inequality in higher dimensions, i.e. given a compact set $K \subset \mathbb{R}^n$, for $Q \in \partial\Omega \cap K$ and $r \in (0, R_K)$ where $R_K > 0$ depends on K there exits a constant $C > 0$ depending on the NTA constants such that
$$\omega^+(B(Q,r)) \cdot \omega^-(B(Q,r)) \leq Cr^{2(n-1)}.$$

2. **Classification of the tangent measures to $\boldsymbol{\omega^{\pm}}$:** this is possible by means of a blow-up procedure compatible with the *singularities* of $u^{\pm}$, the corresponding Green's functions associated to $\omega^{\pm}$. This procedure simultaneously blows up the domains, their boundaries, the harmonic measures and the corresponding Green's functions in such a way that the sub-sequential limits are the harmonic measures $\omega^{\pm}_{\infty}$ and the corresponding Green's functions with pole at infinity $u^{\pm}_{\infty}$ of the blow-up domains $\Omega^{\pm}_{\infty}$ ([KT4]). This blow-up procedure has the additional property that it lifts the singularities of $\nabla u^{\pm}$ at the limit. A remarkable feature is that the 2-sided locally NTA assumption ensures that at points where ω^+ and ω^- are mutually absolutely continuous and the Radon-Nikodym derivative of ω^- with respect to ω^+ is well behaved, all blow-up limits $u_{\infty} = u^+_{\infty} - u^-_{\infty}$ are harmonic polynomials.

3. **Connectivity property of the cone of tangent measures to $\boldsymbol{\omega^{\pm}}$:** this resembles the one used by Preiss in his deep work concerning the rectifiability of measures (see [P]). In our case, the connectivity is a consequence of the fact that u_{∞} is a harmonic polynomial.

Recently Badger (see [Ba]) proved that when Ω is a 2-sided locally NTA domain with ω^+ and ω^- mutually absolutely continuous and the Radon-Nikodym derivative of ω^- with respect to ω^+ is sufficiently regular, $\partial\Omega$ can be decomposed as a finite union of sets Γ_k such that all blow-ups of $\partial\Omega$ at points in Γ_k are the zero set of a homogeneous harmonic polynomial of degree k. More precisely, if Ω is a 2-sided locally NTA domain with ω^+ and ω^- mutually absolutely continuous and $\log f \in VMO(d\omega^+)$ where $f = d\omega^-/d\omega^+$, then there exists $d \in \mathbb{N}$ depending on the NTA constants of Ω such that $\partial\Omega = \Gamma_1 \cup \cdots \cup \Gamma_d$. For $Q \in \Gamma_k$, $1 \leq k \leq d$, all blow ups u_{∞} are homogeneous harmonic polynomials of degree k. Furthermore $\omega^{\pm}(\partial\Omega \backslash \Gamma_1) = 0$.

Question 3.1. Regularity of the set of mutual absolute continuity for $\boldsymbol{\omega^{\pm}}$.

In [B], Bishop asked whether in the case of $\mathbb{R}^n$, $n \geq 3$, if ω^-, ω^+ are mutually absolutely continuous on a set $E \subset \partial\Omega$ and $\omega^{\pm}(E) > 0$, then $\omega^{\pm}$ are mutually absolutely continuous with respect to $\mathcal{H}^{n-1}$ on E. In this case is E $(n-1)$-rectifiable?

The work described above proves that if Ω is a 2-sided locally NTA domain then $\dim_{\mathcal{H}}(E) = n-1$. It also answers Bishop's question under the additional assumption that $\mathcal{H}^{n-1} \llcorner \partial\Omega$ is a Radon measure. In this case E is $(n-1)$-rectifiable. The general question is still open.

Question 3.2. Find a larger class of domains for which Theorem 3.1 holds.

One of the major questions in this area is whether the 2-sided locally NTA assumption can be removed from the hypothesis of Theorem 3.1. In particular it would be very interesting to know to what extent the decomposition theorem holds for John domains (see [A]).

References

[ACF] H. W. Alt, L. A. Caffarelli & A. Friedman, Variational problems with two phases and their free boundaries, *Trans. Amer. Math. Soc.* **282** (1984), 431–461.

[A] A. Ancona, Thèorie du potentiel et domaines de John, *Publicacions Matemtiques* **51** (2007), 345–396.

[Ba] M. Badger, Harmonic polynomials and tangent measures of harmonic measure, preprint. arXiv:0910.2591

[B] C. Bishop, Some questions concerning harmonic measure, in Partial Differential Equations with minimal smoothness and applications, The IMA Volumes in Mathematics and its Applications, Volume 42, 89–98, 1992, edited by Dahlberg, Fabes, R. Fefferman, Jerison, Kenig and Pipher.

[BCGJ] C. Bishop, L. Carleson, J. Garnett & P. Jones, Harmonic measure supported on curves, *Pacific J. Math.*, **138** (1989) 233–236

[CFK] L. Caffarelli, E. Fabes & C. Kenig, Completely singular elliptic-harmonic measures, *Indiana Univ. Math. J.* **30** (1981), 917–924.

[C] L. Carleson, On the support of harmonic functions on sets of Cantor type, *Ann. Acad. Sci. Fenn.*, **10** (1985) 113–123.

[CMS] R. Coifman, Y. Meyer & E. Stein, Some new function spaces and their applications to harmonic analysis, *J. Funct. Anal.* **62** (1985), 304–335.

[D1] B. Dahlberg, Estimates of harmonic measure, *Arch. Rational Mech. Anal.* **65** (1977), no. 3, 275–288.

[D2] B. Dahlberg, On the absolute continuity of elliptic measure, *American Journal of Mathematics* **108** (1986), 1119–1138.

[DJ] G. David & D. Jerison, Lipschitz Approximation to Hypersurfaces, Harmonic Measure, and Singular Integrals, *Indiana Univ. Math. J.* **39** (1990), 831–845.

[DS] G. David & S. Semmes, Singular Integrals and rectifiable sets in $\mathbb{R}^n$. Au-delà des graphes lipschitziens *Astérisque* **193** (1991), 170 pages.

[DKP] M. Dindos, C. Kenig & J. Pipher, *BMO solvability and the A_∞ condition for elliptic operators*, to appear J. Geometric Analysis.

[E] L. Escauriaza, The L^p Dirichlet problem for small perturbations of the Laplacian, *Israel J. Math.* **94** (1996), 353–366.

[EG] L. C. Evans & R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, 1992.

[F] R. Fefferman, A criterion for the absolute continuity of the harmonic measure associated with an elliptic operator, *J. Amer. Math. Soc.* **2** (1989), no. 1, 127–135.

[FKP] R. Fefferman, C. Kenig & J. Pipher, The theory of weights and the Dirichlet problem for elliptic equations, *Ann. of Math.* **134** (1991), 65–124.

[GM] J. Garnett & D. Marshall, *Harmonic Measure*, New Mathematical Monographs, Cambridge University Press, 2005.

[HMT] S. Hofmann, M. Mitrea & M. Taylor, Singular integrals and elliptic boundary problems on regular Semmes-Kenig-Toro domains, to appear in International Mathematics Research Notices, Oxford University Press.

[JK] D. Jerison & C. Kenig, Boundary behavior of harmonic functions in nontangentially accessible domains, *Adv. in Math.* **46** (1982), 80–147.

[JW] P. Jones & T. Wolff, Hausdorff dimension of harmonic measure in the plane, *Acta Math.* **161** (1988), 131–144.

[KP] C. Kenig & J. Pipher, The Neumann problem for elliptic equations with non-smooth coefficients, *Invent. math.* **113** (1993), 447–509.

[KPT] C. Kenig, D. Preiss & T. Toro, Boundary structure and size in terms of interior and exterior harmonic measures in higher dimensions, *J. Amer. Math. Soc.* **22** (2009), 771–796.

[KT1] C. Kenig & T. Toro, Harmonic measure on locally flat domains, *Duke Math. J.* **87** (1997), no. 3, 509–551.

[KT2] C. Kenig & T. Toro, Free Boundary Regularity for harmonic measures and Poisson kernels, *Ann. of Math.* **150** (1999), 369–454

[KT3] C. Kenig & T. Toro, Poisson kernel characterization of Reifenberg flat chord arc domains, *Ann. Scient. Ec. Norm. Sup.* **36** (2003), 323–401.

[KT4] C. Kenig & T. Toro, Free boundary regularity below the continuous threshold: 2-phase problems, *J. Reine Angew. Math.* **596** (2006), 1–44.

[LVV] J. Lewis, G. C. Verchota, & A. Vogel, On Wolff snowflakes, *Pacific J. Math*, **218** (2005), 139–166.

[Mak] N. Makarov, Distortion of boundary sets under conformal mappings, *Proc. London Math. Soc.*, **51** (1985), 369–384.

[MPT] E. Milakis, J. Pipher & T. Toro, Harmonic analysis on chord arc domains, in preparation.

[MM] L. Modica & S. Mortola, Construction of a singular elliptic-harmonic measure, *Manuscripta Math.* **33** (1980), 81–98.

[P] D. Preiss, Geometry of measures in $\mathbb{R}^n$: distribution, rectifiability, and densities, *Ann. of Math.* **125** (1987), 537–643.

[S1] S. Semmes, Analysis vs. Geometry on a Class of Rectifiable Hypersurfaces, *Indiana Univ. J.* **39** (1990), 1005–1035.

[S2] S. Semmes, Chord-Arc Surfaces with Small Constant, II: Good Parametrizations, *Adv. in Math.* **88** (1991), 170–199.

[W] T. Wolff, Counterexamples with harmonic gradients in $\mathbb{R}^3$, *Essays in honor of Elias M. Stein,* Princeton Mathematical Series **42** (1995), 321–384.

Section 9

Functional Analysis and Applications

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Orbit Equivalence and Measured Group Theory

Damien Gaboriau*

Abstract

We give a survey of various recent developments in orbit equivalence and measured group theory. This subject aims at studying infinite countable groups through their measure preserving actions.

Mathematics Subject Classification (2000). Primary 37A20; Secondary 46L10.

Keywords. Orbit equivalence, Measured group theory, von Neumann algebras

1. Introduction

Orbit equivalence and measure equivalence theories deal with countable groups Γ acting on standard measure spaces and with the associated orbit partitions of the spaces. This is very much connected from its birth with operator algebras [MvN36]; many of the recent progresses in both areas were made conjointly (see [Pop07b, Vae07, Vae10]). It turns out to be also connected with geometric group theory (see section 9 and [Fur09]), descriptive set theory (see [JKL02, KM04]), percolation on graphs (see [LP09])... with fruitful cross-pollination.

There are many examples of mathematical domains where the orbit equivalence or measured approach helps solving delicate questions involving countable groups Γ. For instance, in connection with group ℓ^2-Betti numbers $\beta_n^{(2)}(\Gamma)$, this was useful to attack:

- various vanishing results in [Gab02, ST07];
- the study of harmonic Dirichlet functions on percolation subgraphs [Gab05b];
- the comparison between the uniform isoperimetric constant and $\beta_1^{(2)}(\Gamma)$ [LPV08];

*Unité de Mathématiques Pures et Appliquées, Université de Lyon, CNRS, ENS Lyon, 69364 Lyon cedex 7, FRANCE. E-mail: damien.gaboriau@ens-lyon.fr.

– problems of topological nature, related to the work of Gromov about the minimal volume [Sau09].

In geometric group theory, the quasi-isometry invariance of various cohomological properties for amenable groups [Sha04] was obtained that way. Gaboriau-Lyons' measurable solution to von Neumann's problem [GL09] happens to be a way to extend results about groups containing a copy of the free group $\mathbf{F}_2$ to every non-amenable group (see section 10). This was used by [Eps08] and in Dixmier's unitarizability problem [EM09, MO10].

The purpose of this survey is to describe some foundations of the theory and some of its most recent developments. There are many aspects upon which we shall inevitably not touch here, and many results are just alluded to, with as far as possible the relevant bibliography.

There are several excellent books and surveys with various focuses on orbit equivalence to which the reader is referred for further information, for instance [KM04, Gab05a, Sha05, Pop07b, Fur09, Kec10].

2. Setting and Examples

The measure spaces X will always be assumed to be standard Borel spaces and unless specified otherwise, the measure μ will be a non-atomic probability measure. Measurably, (X, μ) is isomorphic to the interval $([0, 1], \mathrm{Leb})$ equipped with the Lebesgue measure. Moreover, the actions $\Gamma \curvearrowright^{\alpha} (X, \mu)$ we shall consider will be by Borel automorphisms and probability measure preserving (p.m.p.), i.e. $\forall \gamma \in \Gamma$, $A \subset X$: $\mu(\gamma.A) = \mu(A)$ (we only consider Borel sets). Shortly, α is a **p.m.p. action** of Γ. In this measured context, null sets are neglected. Equality for instance is always understood almost everywhere. The action α is **(essentially) free** if $\mu\{x : \ \gamma.x = x\} > 0 \Rightarrow \gamma = id$. The action is **ergodic** if the dynamics is indecomposable, i.e. whenever X admits a partition $X = A \cup {}^c A$ into invariant Borel subsets, then one of them is trivial, i.e. $\mu(A)\mu({}^c A) = 0$.

We now present a series of basic examples which shall already exhibit a rich variety of phenomena.

Example 2.1. The action of $\mathbb{Z}^n$ on the circle $\mathbb{S}^1$ by rationally independent rotations.

Example 2.2. The standard action $\mathrm{SL}(n, \mathbb{Z}) \curvearrowright \mathbb{T}^n$ on the n-torus $\mathbb{R}^n/\mathbb{Z}^n$ with the Lebesgue measure. The behavior is drastically different for $n \geq 3$ and for $n = 2$. The higher dimensional case was central in the super-rigidity results of Zimmer [Zim84] and Furman [Fur99a, Fur99b] (see section 12). The 2-dimensional case $\mathrm{SL}(2, \mathbb{Z}) \curvearrowright \mathbb{T}^2$ played a particularly important role in the recent developments of the theory, mainly because of its relation with the semi-direct product $\mathrm{SL}(2, \mathbb{Z}) \ltimes \mathbb{Z}^2$, in which $\mathbb{Z}^2$ is known to have the so called relative property (T) (see section 11), while $\mathrm{SL}(2, \mathbb{Z})$ is a virtually free group (it has a finite index free subgroup).

Example 2.3. Volume-preserving group actions on finite volume manifolds.

Example 2.4. Given two lattices Γ, Λ in a Lie group H (or more generally a locally compact second countable group) the actions by left (resp. right by the inverse) multiplication on H induce actions on the finite measure standard spaces $\Gamma \curvearrowright H/\Lambda$ and $\Lambda \curvearrowright \Gamma\backslash H$.

Example 2.5. A compact group K, its Haar measure μ and the action of a countable subgroup Γ by left multiplication on K.

Example 2.6. Let (X_0, μ_0) be a standard probability measure space, possibly with atoms[1]. The **standard Bernoulli shift action** of Γ is the action on the space X^Γ of sequences $(x_\gamma)_{\gamma\in\Gamma}$ by shifting the indices $g.(x_\gamma)_{\gamma\in\Gamma} = (x_{g^{-1}\gamma})_{\gamma\in\Gamma}$, together with the Γ-invariant product probability measure $\otimes_\Gamma \mu_0$. In particular, every countable group admits at least one p.m.p. action. The action is free (and ergodic) iff Γ is infinite.

More generally, consider some action $\Gamma \curvearrowright \mathtt{V}$ of Γ on some countable set $\mathtt{V}$. The **generalized Bernoulli shift action** of Γ is the action on the space $X^{\mathtt{V}}$ of sequences $(x_v)_{v\in\mathtt{V}}$ by shifting the indices $g.(x_v)_{v\in\mathtt{V}} = (x_{g^{-1}.v})_{v\in\mathtt{V}}$, with the invariant product probability measure.

Example 2.7. Profinite actions. Consider an action $\Gamma \curvearrowright (\mathtt{T}, v_0)$ of Γ on a locally finite rooted tree. The action preserves the equiprobability on the levels, and the induced limit probability measure on the set of ends of the tree is Γ-invariant. For instance, if Γ is residually finite, as witnessed by a chain of finite index subgroups $\Gamma = \Gamma_0 > \Gamma_1 > \Gamma_2 > \cdots \Gamma_i > \cdots$ with trivial intersection, such a rooted tree $(\mathtt{T}, (v_0 = \Gamma/\Gamma_0))$ is naturally built with vertex set (of level i) the cosets Γ/Γ_i and edges given by the reduction maps $\Gamma/\Gamma_{i+1} \to \Gamma/\Gamma_i$. The action is ergodic iff it is transitive on the levels.

A first connection with functional analysis is made through the following. The **Koopman representation** of a p.m.p. action $\Gamma\curvearrowright^\alpha (X, \mu)$ is the representation κ_α of Γ on $L^2_0(X, \mu)$ given by[2] $\kappa_\alpha(\gamma)(\xi)(x) = \xi(\alpha(\gamma^{-1})(x))$ [Koo31].

A lot of dynamical properties of the action are read from this unitary representation and its spectral properties. For instance, the action is ergodic if and only if its Koopman representation has no Γ-invariant unit vector. In examples 2.1 and 2.2 or 2.6, various properties are deduced from the fact that the Koopman representation admits a Hilbert basis which is either made of eigenvectors or permuted by Γ (see for instance [Sch80], [KT08]). The classical ergodic theory considers such actions up to **conjugacy** (**notation:** $\Gamma_1\curvearrowright^{\alpha_1} X_1 \overset{\text{Conj}}{\sim} \Gamma_2\curvearrowright^{\alpha_2} X_2$).

[1] for instance $X_0 = \{0, 1\}$ and $\mu_0(\{0\}) = 1 - p, \mu_0(\{1\}) = p$ for some $p \in (0, 1)$. The only degenerate situation one wishes to avoid is X_0 consisting of one single atom.

[2] The constants are fixed vectors for the representation on $L^2(X, \mu)$. Its orthocomplement $L^2_0(X, \mu) = L^2(X, \mu)\ominus\mathbb{C}1$ consists of $\{\xi \in L^2(X, \mu) : \int_X \xi(x)d\mu(x) = 0\}$.

We now introduce a weaker notion of equivalence and turn from classical ergodic theory to orbit equivalence theory. Here $\Gamma.x$ denotes the orbit of x under the Γ-action.

Definition 2.8 (Orbit equivalence). Two actions $\Gamma_i \curvearrowright^{\alpha_i} (X_i, \mu_i)$ (for $i = 1, 2$) are **orbit equivalent (OE)** (**notation:** $\Gamma_1 \curvearrowright^{\alpha_1} X_1 \overset{\text{OE}}{\sim} \Gamma_2 \curvearrowright^{\alpha_2} X_2$) if there is a measured space isomorphism[3] $f : X_1 \to X_2$ that sends orbits to orbits:

$$\text{for a.a. } x \in X_1\text{: } f(\Gamma_1.x) = \Gamma_2.f(x).$$

In particular, the groups are no longer assumed to be isomorphic. When studying actions up to orbit equivalence, what one is really interested in, is the partition of the space into orbits or equivalently the **orbit equivalence relation**:

$$\mathcal{R}_\alpha := \{(x, y) \in X : \exists \gamma \in \Gamma \text{ s.t. } \alpha(\gamma)(x) = y\}. \tag{1}$$

This equivalence relation satisfies the following three properties: (1) its classes are (at most) countable, (2) as a subset of $X \times X$, it is measurable, (3) it **preserves the measure** μ: this means that every measurable automorphism $\phi : X \to X$ that is **inner** (x and $\phi(x)$ belong to the same class for a.a. $x \in X$) has to preserve μ.

Axiomatically [FM77a], the object of study is an equivalence relation $\mathcal{R}$ on (X, μ) satisfying the above three conditions: we simply call it a **p.m.p. equivalence relation**. Two p.m.p. equivalence relations $\mathcal{R}_1, \mathcal{R}_2$ will be **orbit equivalent** if there is a measured space isomorphism $f : X_1 \to X_2$ sending classes to classes.

This abstraction is necessary when one wants to consider, for instance, the **restriction** $\mathcal{R}|A$ of $\mathcal{R}$ to some non-null Borel subset $A \subset X$: the standard Borel space A is equipped with the normalized probability measure $\mu_A(C) = \mu(C)/\mu(A)$ and $(x, y) \in \mathcal{R}|A \Leftrightarrow x, y \in A$ and $(x, y) \in \mathcal{R}$.

In fact, this more general context allows for much more algebraic flexibility since the lattice of subrelations of $\mathcal{R}_\alpha$ for some Γ-action α is much richer than that of subgroups of Γ (see von Neumann's problem in section 10). Also, $\mathcal{R}_\alpha$ is easier to decompose as a "free product or a direct product" than Γ itself (see section 7 and [AG10]).

By an **increasing approximation** $\mathcal{R}_n \nearrow \mathcal{R}$ of a p.m.p. equivalence relation $\mathcal{R}$ we mean an increasing sequence of standard (p.m.p.) equivalence subrelations with $\cup_n \mathcal{R}_n = \mathcal{R}$.

An important notion is that of hyperfiniteness: a p.m.p. equivalence relation $\mathcal{R}$ is **hyperfinite** if it admits an increasing approximation by **finite** equivalence subrelations $\mathcal{R}_n$ (i.e. the classes of the $\mathcal{R}_n$ are finite). Obviously all the actions of locally finite groups (i.e. groups all of whose finitely generated subgroups are

[3] An isomorphism of measure spaces is defined almost everywhere and respects the measures: $f_* \mu_1 = \mu_2$.

finite) generate orbit equivalence relations in this class; for instance such groups as $\Gamma = \oplus_{\mathbb{N}}\Lambda_n$, where the Λ_n are finite. This is also the case for all $\mathbb{Z}$-actions. Dye's theorem is among the fundamental theorems in orbit equivalence theory:

Theorem 2.9 ([Dye59]). *All the ergodic hyperfinite p.m.p. equivalence relations are mutually orbit equivalent.*

A series of results due in particular to Dye, Connes, Krieger, Vershik... leads to Ornstein-Weiss' theorem (see [CFW81] for a more general version):

Theorem 2.10 ([OW80]). *If Γ is amenable then all its p.m.p. actions are hyperfinite.*

In particular, when ergodic, these actions are indistinguishable from the orbit equivalence point of view! All the usual ergodic theoretic invariants are lost. This common object will be denoted $\mathcal{R}_{\text{hyp}}$. On the other hand, if Γ admits a free p.m.p. hyperfinite action, then Γ has to be amenable, thus showing the border of this huge singular area that produces essentially a single object. The non-amenable world is much more complicated and richer.

3. The Full Group

The **full group** of $\mathcal{R}$ denoted by $[\mathcal{R}]$ is defined as the group of p.m.p. automorphisms of (X, μ) whose graph is contained in $\mathcal{R}$:

$$[\mathcal{R}] := \{T \in \text{Aut}(X, \mu) : (x, T(x)) \in \mathcal{R} \text{ for a.a. } x \in X\}.$$

It was introduced and studied by Dye [Dye59], and it is clearly an OE-invariant. But conversely, its algebraic structure is rich enough to remember the equivalence relation:

Theorem 3.1 ([Dye63]). *(Dye's reconstruction theorem) Two ergodic p.m.p. equivalence relations $\mathcal{R}_1$ and $\mathcal{R}_2$ are OE iff their full groups are algebraically isomorphic; moreover the isomorphism is then implemented by an orbit equivalence.*

The full group has very nice properties. The topology given by the bi-invariant metric $d(T, S) = \mu\{x : T(x) \neq S(x)\}$ is Polish. In general, it is not locally compact and, in fact, homeomorphic with the separable Hilbert space ℓ^2 [KT10].

Theorem 3.2 ([BG80, Kec10]). *The full group is a simple group iff $\mathcal{R}$ is ergodic.*

And it satisfies this very remarkable, automatic continuity:

Theorem 3.3 (Kittrell-Tsankov [KT10]). *If $\mathcal{R}$ is ergodic, then every group homomorphism $f : [\mathcal{R}] \to G$ with values in a separable topological group is automatically continuous.*

Hyperfiniteness translates into an abstract topological group property:

Theorem 3.4 (Giordano-Pestov [GP07]). *Assuming $\mathcal{R}$ ergodic, $\mathcal{R}$ is hyperfinite iff $[\mathcal{R}]$ is extremely amenable.*

Recall that a topological group G is **extremely amenable** if every continuous action of G on a (Hausdorff) compact space has a fixed point. Together with Kittrell-Tsankov's result, this gives that every action of $[\mathcal{R}_{\text{hyp}}]$ by homeomorphisms on a compact metrizable space has a fixed point.

Closely related to the full group, the **automorphism group** $\text{Aut}(\mathcal{R}) := \{T \in \text{Aut}(X,\mu) : (x,y) \in \mathcal{R} \Rightarrow (T(x),T(y)) \in \mathcal{R} \text{ for a.a. } x \in X\} \triangleright [\mathcal{R}]$ and the **outer automorphism group** (the quotient) $\text{Out}(\mathcal{R}) = \text{Aut}(\mathcal{R})/[\mathcal{R}]$ have attracted much attention for several years; see for instance [GG88a, Gef93, Gef96, Fur05, IPP08, Pop06b, Kec10, Kid08c, PV08d, PV08a, Gab08] and references therein and section 11.

4. Associated von Neumann Algebra

In fact, the original interest for orbit equivalence came from its connection with von Neumann algebras. Murray and von Neumann [MvN36] considered p.m.p. group actions $\Gamma \curvearrowright^{\alpha} (X,\mu)$ as a machine to produce finite von Neumann algebras M_α, via their group-measure-space construction. And Singer [Sin55] was the first to explicitly notice that M_α only depends on the OE class of the action. Feldman-Moore [FM77b] extended the group-measure-space construction to the context of p.m.p. equivalence relations.

A p.m.p. equivalence relation $\mathcal{R}$ on (X,μ), considered as a Borel subspace of $X \times X$ is naturally equipped with a (*a priori* infinite) **measure** ν. It is defined as follows: for every Borel subset $C \subset \mathcal{R}$,

$$\nu(C) = \int_X |\pi_l^{-1}(x) \cap C| d\mu(x), \tag{2}$$

where $\pi_l : \mathcal{R} \to X$ is the projection onto the first coordinate, $\pi_l^{-1}(x)$ is the fiber above $x \in X$, and $|\pi_l^{-1}(x) \cap C|$ is the (at most countable) cardinal of its intersection with C. A similar definition could be made with the projection π_r on the second coordinate instead, but the fact that $\mathcal{R}$ is p.m.p. ensures that these two definitions would coincide.

The (generalized) **group-measure-space von Neumann algebra** $L(\mathcal{R})$ **associated with** $\mathcal{R}$ is generated by two families of operators of the separable Hilbert space $L^2(\mathcal{R},\nu)$: $\{L_g : g \in [\mathcal{R}]\}$ and $\{L_f : f \in L^\infty(X,\mu)\}$, where $L_g\xi(x,y) = \xi(g^{-1}x,y)$ and $L_f\xi(x,y) = f(x)\xi(x,y)$ for every $\xi \in L^2(\mathcal{R},\nu)$. It contains $\{L_f : f \in L^\infty(X,\mu)\} \simeq L^\infty(X,\mu)$ as a **Cartan subalgebra** (i.e. a maximal abelian subalgebra whose normalizer generates $L(\mathcal{R})$). With this definition, $L(\mathcal{R})$ is clearly an OE-invariant.

Definition 4.1 (von Neumann equivalence or W*-equivalence). Two p.m.p. equivalence relations $\mathcal{R}_i$ on (X_i, μ_i) (for $i = 1, 2$) are **von Neumann equivalent** or **W*-equivalent** if $L(\mathcal{R}_1) \simeq L(\mathcal{R}_2)$ (**notation:** $\mathcal{R}_1 \overset{\text{vN}}{\sim} \mathcal{R}_2$).

There exist non-OE equivalence relations producing isomorphic $L(\mathcal{R})$ ([CJ82], [OP08b]). Indeed, the additionnal data needed to recover $\mathcal{R}$ is the embedding $L^\infty(X, \mu) \subset L(\mathcal{R})$ of the Cartan subalgebra inside $L(\mathcal{R})$ (up to isomorphisms) [Sin55, FM77b].

5. Strong Ergodicity

Recall that a standard p.m.p. equivalence relation $\mathcal{R}$ is **ergodic** if every $\mathcal{R}$-invariant[4] Borel set $A \subset X$ satisfies $\mu(A)(\mu(A) - 1) = 0$. The notion of strong ergodicity was introduced by Schmidt as an OE-invariant.

Definition 5.1 ([Sch80]). An ergodic p.m.p. countable standard equivalence relation $\mathcal{R}$ is **strongly ergodic** if every almost invariant sequence[5] of Borel subsets $A_n \subset X$ is trivial, i.e. satisfies $\lim_{n\to\infty} \mu(A_n)(1 - \mu(A_n)) = 0$.

There are several equivalent definitions of strong ergodicity, see for instance [JS87]. We give yet another one below through approximations.

Proposition 5.2. *An ergodic equivalence relation $\mathcal{R}$ is **strongly ergodic** if and only if every increasing approximation $\mathcal{R}_n \nearrow \mathcal{R}$ admits an ergodic restriction $\mathcal{R}_n|U$ to some non-negligeable Borel set U, for big enough n.*

In other words, for big enough n the ergodic decomposition of $\mathcal{R}_n$ admits an atom. It is easy to see that whenever a p.m.p. action $\Gamma \curvearrowright (X, \mu)$ is non-strongly ergodic, its Koopman representation κ_0 almost has invariant vectors. The converse does not hold in general [Sch81], [HK05]. However, Chifan-Ioana [CI10] extending an argument of Abert-Nikolov [AN07] proved that this is indeed the case when the commutant of $\Gamma \curvearrowright (X, \mu)$ in $\mathrm{Aut}(X, \mu)$ acts ergodically on (X, μ). Standard Bernoulli shifts are strongly ergodic iff the group is non-amenable. In particular every non-amenable group admits at least one strongly ergodic action.

Kechris-Tsankov [KT08] characterized the generalized Bernoulli shifts $\Gamma \curvearrowright (X_0, \mu_0)^{\mathtt{V}}$ that are strongly ergodic as those for which the action $\Gamma \curvearrowright \mathtt{V}$ is **non-amenable** (i.e. the representation on $\ell^2(\mathtt{V})$ does not admit any sequence of almost invariant vectors).

The consideration of the Koopman representation κ_0 ensures that for (infinite) groups with Kazhdan property (T) **every** ergodic p.m.p. action is strongly ergodic. And Connes-Weiss (by using Gaussian random variables) showed that this is a criterion for property (T) [CW80].

[4] for each g in the full group $[\mathcal{R}]$, $\mu(A \Delta gA) = 0$.
[5] i.e. for each g in $[\mathcal{R}]$, $\lim_{n\to\infty} \mu(A_n \Delta gA_n) = 0$.

A graphing Φ (see section 6) on X naturally defines a "metric" d_Φ on X: the simplicial distance associated with the graph structure in the classes of $\mathcal{R}_\Phi$ and $d_\Phi = \infty$ between two points in different classes. This is a typical instance of what Gromov calls a ***mm*-space** [Gro00], i.e. a probability measure space (X, μ) together with a Borel function $d : X \times X \to \mathbb{R}^+ \cup \{\infty\}$ satisfying the standard metric axioms except that one allows $d(x, x') = \infty$. A mm-space (X, μ, d) is **concentrated** if $\forall \delta > 0$, there is $\infty > r_\delta > 0$ such that $\mu(A), \mu(B) \geq \delta \Rightarrow d(A, B) \leq r_\delta$. For instance, if $\Phi = \langle \varphi_1 : X \to X \rangle$ is given by a single p.m.p. ergodic isomorphism, (X, μ, d_Φ) is never concentrated. Gromov observed for finitely generated groups that every p.m.p. ergodic action of Γ has (respectively, never has) the concentration property if Γ has Kazhdan's property (T) (respectively, if Γ is amenable). Pichot made the connection with strong ergodicity:

Theorem 5.3 ([Pic07a]). *Let $\Phi = (\varphi_i)_{i=1,\cdots,p}$ be a graphing made of finitely many partial isomorphisms. The space (X, μ, d_Φ) is concentrated iff $\mathcal{R}_\Phi$ is strongly ergodic.*

See also [Pic07b] for a characterization of strong ergodicity (as well as of property (T) or amenability) in terms of the spectrum of diffusion operators associated with random walks on the equivalence relation $\mathcal{R}$.

For the standard $\mathrm{SL}(2, \mathbb{Z})$ action on the 2-torus $\mathbb{R}^2/\mathbb{Z}^2$, every non-amenable subgroup $\Lambda < \mathrm{SL}(2, \mathbb{Z})$ acts ergodically, and even strongly ergodically. Similarly for the generalized Bernoulli shift $\Gamma \curvearrowright (X_0, \mu_0)^{\mathtt{V}}$, where the stabilizers of the action $\Gamma \curvearrowright \mathtt{V}$ are amenable. Inspired by [CI10], define more generally:

Definition 5.4 (Solid ergodicity). A p.m.p. standard equivalence relation $\mathcal{R}$ is called **solidly ergodic** if for every (standard) subrelation $\mathcal{S}$ there exists a measurable partition $\{X_i\}_{i \geq 0}$ of X in $\mathcal{S}$-invariant subsets such that:

(a) the restriction $\mathcal{S}|X_0$ is hyperfinite

(b) the restrictions $\mathcal{S}|X_i$ are strongly ergodic for every $i > 0$.

In particular, an ergodic subrelation of a solidly ergodic relation is either hyperfinite or strongly ergodic. By Zimmer [Zim84, Prop. 9.3.2], every ergodic p.m.p. standard equivalence relation $\mathcal{R}$ contains an ergodic hyperfinite subrelation $\mathcal{S}$ which, being non strongly ergodic, contains an aperiodic subrelation with diffuse ergodic decomposition. Thus the X_0 part cannot be avoided, even for aperiodic subrelations.

One gets an equivalent definition if one replaces "strongly ergodic" by "ergodic" (see [CI10, Prop. 6] for more equivalent definitions). It may seem quite unlikely that such relations really exist. However, Chifan-Ioana [CI10] observed that the notion of solidity and its relative versions introduced by Ozawa [Oza04] (by playing between C*- and von Neumann algebras) imply solid ergodicity (hence the name). Moreover, they established a general solidity result for Bernoulli shifts.

Theorem 5.5. *The following actions are solidly ergodic:*

- *The standard action* $\mathrm{SL}(2,\mathbb{Z}) \curvearrowright \mathbb{R}^2/\mathbb{Z}^2$ *[Oza09].*
- *The generalized Bernoulli action* $\Gamma \curvearrowright (X_0,\mu_0)^V$*, when the* Γ*-action* $\Gamma \curvearrowright V$ *has amenable stabilizers [CI10].*

When the group Γ is exact[6], the above statement for the standard Bernoulli shifts also follows from [Oza06, Th. 4.7].

A positive answer to the following percolation-theoretic question would give another proof of solid ergodicity for the standard Bernoulli shifts:

Question 5.6. Let Γ be a countable group with a finite generating set S. Let $\pi : (X_0,\mu_0)^\Gamma \to [0,1]$ be any measure preserving map (i.e. $\pi_*(\otimes_\Gamma \mu_0) = \mathrm{Leb}$) and Φ_π be the "fiber-graphing" made of the restriction φ_s of $s \in S$ to the set $\{\omega \in (X_0,\mu_0)^\Gamma : \pi(s.\omega) = \pi(\omega)\}$. Is the equivalence relation generated by Φ_π finite?

6. Graphings

The **cost** of a p.m.p. equivalence relation $\mathcal{R}$ has been introduced by Levitt [Lev95]. It has been studied intensively in [Gab98, Gab00a]. See also [KM04, Kec10, Fur09] and the popularization paper [Gab10b]. When an equivalence relation is generated by a group action, the relations between the generators of the group introduce redundancy in the generation, and one can decrease this redundancy by using instead **partially defined isomorphisms**.

A countable family $\Phi = (\varphi_j : A_j \xrightarrow{\sim} B_j)_{j\in J}$ of measure preserving isomorphisms between Borel subsets $A_i, B_i \subset X$ is called a **graphing**. It generates a p.m.p. equivalence relation $\mathcal{R}_\Phi$: the smallest equivalence relation such that $x \sim \varphi_j(x)$ for $j \in J$ and $x \in A_j$. Moreover, Φ furnishes a graph structure (hence the name) $\Phi[x]$ on the class of each point $x \in X$: two points y and z in its class are connected by an edge whenever $z = \varphi_j^{\pm 1}(y)$ for some $j \in J$. If $\mathcal{R}$ is generated by a free action of Γ and if Φ is made of isomorphisms associated with a generating set S of Γ, then the graphs $\Phi[x]$ are isomorphic with the corresponding Cayley graph of Γ. When all the graphs $\Phi[x]$ are trees, Φ is called a **treeing**. If it admits a generating treeing, $\mathcal{R}$ is called **treeable**. See Adams [Ada88, Ada90] for the first study of treed equivalence relations.

The **cost** of Φ is the number of generators weighted by the measure of their support: $\mathrm{Cost}(\Phi) = \sum_{j\in J} \mu(A_j) = \sum_{j\in J} \mu(B_j)$. The **cost of** $\mathcal{R}$ is the infimum over the costs of its generating graphings: $\mathrm{Cost}(\mathcal{R}) = \inf\{\mathrm{Cost}(\Phi) : \mathcal{R} = \mathcal{R}_\Phi\}$. It is by definition an OE-invariant. The cost of $\mathcal{R}$ is ≥ 1 when the classes are infinite [Lev95]. Together with Ornstein-Weiss' theorem this

[6]Recall that a discrete group Γ is **exact** iff it acts amenably on some compact topological space.

gives that every p.m.p. free action of an infinite amenable group has cost = 1. Various commutation properties in a group Γ also entail cost = 1 for all of its free actions. For instance when $\Gamma = G \times H$ is the product of two infinite groups and contains at least one infinite order element or $\Gamma = \mathrm{SL}(n, \mathbb{Z})$, for $n \geq 3$. It is not difficult to see that when a finite cost graphing Φ realizes the cost of $\mathcal{R}_\Phi$ then Φ is a treeing. The main results in [Gab98] claim the converse:

Theorem 6.1. *If Φ is a treeing then $\mathrm{Cost}(\mathcal{R}_\Phi) = \mathrm{Cost}(\Phi)$. In particular, the free actions of the free group $\mathbf{F}_n$ have cost n.*

In particular, free groups of different ranks cannot have OE free actions. The cost measures the amount of information needed to construct $\mathcal{R}$. It is an analogue of the **rank** of a countable group Γ, i.e. the minimal number of generators or in a somewhat pedantic formulation, the infimum of the measures $\delta(S)$ over the generating systems S, where δ denotes the counting measure on the group. Similarly the **cost of** $\mathcal{R}$ is the infimum of the measures $\nu(C)$ over the Borel subsets $C \subset \mathcal{R}$ which generate $\mathcal{R}$, where ν is the measure on $\mathcal{R}$ introduced in section 4, equation (2) (compare Connes' Bourbaki seminar [Con04]).

In [Gab00a] the notion of **free product decomposition** $\mathcal{R} = \mathcal{R}_1 * \mathcal{R}_2$ (and more generally **free product with amalgamation** $\mathcal{R} = \mathcal{R}_1 *_{\mathcal{R}_3} \mathcal{R}_2$) of an equivalence relation over subrelations is introduced (see also [Ghy95, Pau99]). Of course, when $\mathcal{R}$ is generated by a free action of a group, a decomposition of $\Gamma = \Gamma_1 *_{\Gamma_3} \Gamma_2$ induces the analogous decomposition of $\mathcal{R} = \mathcal{R}_{\Gamma_1} *_{\mathcal{R}_{\Gamma_3}} \mathcal{R}_{\Gamma_2}$. The cornerstone in cost theory is the following computation:

Theorem 6.2 ([Gab00a]). $\mathrm{Cost}(\mathcal{R}_1 *_{\mathcal{R}_3} \mathcal{R}_2) = \mathrm{Cost}(\mathcal{R}_1) + \mathrm{Cost}(\mathcal{R}_2) - \mathrm{Cost}(\mathcal{R}_3)$, *when $\mathcal{R}_3$ is hyperfinite (possibly trivial).*

These techniques allow for the calculation of the cost of the free actions of several groups: for instance $\mathrm{SL}(2, \mathbb{Z})$ (Cost $= 1 + 1/12$), surface groups $\pi_1(\Sigma_g)$ (Cost $= 2g-1$)... In all the examples computed so far, the cost does not depend on the particular free action of the group, thus raising the following question (which proved to be related to **rank gradient** and a low-dimensional topology problem; see [AN07]) (see also Question 8.2):

Question 6.3 (Fixed Price Problem). Does there exist a group Γ with two p.m.p. free actions of non equal costs?

Observe that both the infimum $\mathrm{Cost}(\Gamma)$ ([Gab00a]) and the supremum $\mathrm{Cost}^*(\Gamma)$ ([AW]) among the costs of all free p.m.p. actions of Γ are realized by some actions.

Question 6.4 (Cost for Kazdhan groups). Does there exist a Kazdhan property (T) group with a p.m.p. free action of cost > 1?

In his very rich monograph [Kec10], Kechris studied the continuity properties of the cost function on the space of actions and proved that $\mathrm{Cost}(\mathcal{R}) > 1$ for an ergodic $\mathcal{R}$ forces its outer automorphism group to be Polish. He also

introduced the topological OE-invariant $t([\mathcal{R}])$, defined as the minimum number of generators of a dense subgroup of the full group $[\mathcal{R}]$ and related it with the cost [Kec10]. When $\mathcal{R}$ is generated by a free ergodic action of $\mathbf{F}_n$, Miller obtained the following lower bound: $n+1 \leq t([\mathcal{R}])$, and [KT10] proved that $t([\mathcal{R}_{\text{hyp}}]) \leq 3$ and that $t([\mathcal{R}]) \leq 3(n+1)$.

Lyons-Pichot-Vassout [LPV08] introduced the **uniform isoperimetric constant** $h(\mathcal{R})$ for p.m.p. equivalence relations, a notion similar to that for countable groups $h(\Gamma)$. They were able to obtain the purely group theoretic sharp comparison $2\beta_1^{(2)}(\Gamma) \leq h(\Gamma)$ (where $\beta_1^{(2)}(\Gamma)$ is the first ℓ^2-Betti number). Two complementary inequalities from [LPV08, PV09a] lead to "$2(\text{Cost}(\mathcal{R}) - 1) = h(\mathcal{R})$", thus identifying two OE-invariants of apparently different nature. See [LP09] for an application of cost to percolation theory.

7. Dimensions

Geometric group theory studies countable groups through their actions on "nice spaces". Similarly, for a p.m.p. equivalence relation (it is a groupoid [ADR00]) $\mathcal{R}$ on (X, μ), one might consider its **actions** on fields of spaces $X \ni x \mapsto \Sigma_x$, or **$\mathcal{R}$-field**. For instance, a graphing Φ defines a **measurable field of graphs** $x \mapsto \Phi[x]$, on which the natural isomorphism $\Phi[y] \simeq \Phi[z]$ for $(y, z) \in \mathcal{R}_\Phi$ induces an action of $\mathcal{R}_\Phi$. The Bass-Serre theory [Bas76, Ser77] relates the actions of a group on trees to its free product with amalgamation decompositions (and HNN-extensions). Alvarez [Alv09b, Alv09a] developped an analogous theory in the framework of equivalence relations. For instance an equivalence relation $\mathcal{R}$ acts "properly" on a field of trees iff $\mathcal{R}$ is treeable [Alv09b]. He also obtained a theorem describing the structure of subrelations of a free product [Alv09a], analogous to Kurosh's theorem. This led in [AG10] to the essential uniqueness of a free product decomposition $\mathcal{R} = \mathcal{R}_1 * \cdots * \mathcal{R}_n$ when the factors are **freely indecomposable** (i.e. indecomposable as a non-trivial free product) (compare [IPP08, CH10]). See also [Sak09b] for similar results for some free products with amalgamation over amenable groups.

Definition 7.1 ([AG10]). A countable group is called **measurably freely indecomposable** ($\mathcal{MFI}$) if all its free p.m.p. actions are freely indecomposable.

Examples of $\mathcal{MFI}$ groups are provided by non-amenable groups with $\beta_1^{(2)} = 0$.

Question 7.2 ([AG10]). Produce a $\mathcal{MFI}$ group with $\beta_1^{(2)} > 0$.

More generally, a **simplicial $\mathcal{R}$-field** is a measurable field of simplicial complexes with a simplicial action of $\mathcal{R}$ (see [Gab02]): the space $\Sigma^{(0)}$ of 0-cells has a Borel structure and a measurable map π onto X with countable fibers. The cells are defined in the fibers; $\mathcal{R}$ permutes the fibers; and everything is measurable. The action is **discrete** (or **smooth**, or **proper**) if it admits a

measurable fundamental domain in $\Sigma^{(0)}$. For example, consider a free p.m.p. action $\Gamma \curvearrowright^{\alpha} (X, \mu)$ and a free action of Γ on a (usual, countable) simplicial complex L. This defines a proper simplicial $\mathcal{R}_\alpha$-action on $X \times L$ induced by the diagonal Γ-action. It is instructive to consider an OE action $\Lambda \curvearrowright^{\beta} (X, \mu)$ and to try to figure out the action of $\mathcal{R}_\beta = \mathcal{R}_\alpha$ on $X \times L$ once Γ is forgotten.

The **geometric dimension** geo-dim($\mathcal{R}$) of $\mathcal{R}$ is defined as the smallest possible dimension of a *proper* $\mathcal{R}$-field of *contractible* simplicial complexes [Gab02]. It is analogous to (and bounded above by) the classical geometric dimension ([Bro82]) of Γ. The **approximate dimension** [Gab02] (no classical analogue) approx-dim($\mathcal{R}$) of $\mathcal{R}$ is defined as the smallest possible upper limit of geometric dimensions along increasing approximations of $\mathcal{R}$:

$$\text{approx-dim}(\mathcal{R}) := \min\{\sup(\text{geo-dim}(\mathcal{R}_n))_n : (\mathcal{R}_n) \nearrow \mathcal{R}\}.$$

For instance, geo-dim($\mathcal{R}$) = 0 for finite equivalence relations; approx-dim($\mathcal{R}$) = 0 iff $\mathcal{R}$ is hyperfinite; and geo-dim($\mathcal{R}$) = 1 iff $\mathcal{R}$ is treeable. Thus, quite surprisingly, surface groups admit free actions of geo-dim = 1. Every free action of a Kazhdan property (T) group satisfies approx-dim = geo-dim > 1 [AS90, Moo82, Gab10a]. In the following statement, $\beta_n^{(2)}$ denotes the n-th ℓ^2-Betti number (see section 8).

Theorem 7.3 ([Gab10a]). *These dimensions satisfy:*
-a- geo-dim($\mathcal{R}$) − 1 ≤ approx-dim($\mathcal{R}$) ≤ geo-dim($\mathcal{R}$).
-b- If $\Lambda < \Gamma$ satisfies $\beta_p^{(2)}(\Lambda) \neq 0$, then geo-dim($\mathcal{R}_\alpha$) ≥ p for every free p.m.p. action $\Gamma \curvearrowright^{\alpha} (X, \mu)$. If moreover geo-dim($\mathcal{R}_\alpha$) = p, then $\beta_p^{(2)}(\Gamma) \neq 0$.

It follows that every free action of $\mathbf{F}_{r_1} \times \cdots \times \mathbf{F}_{r_p}$ ($r_j \geq 2$) (resp. $\mathbb{Z} \times \mathbf{F}_{r_1} \times \cdots \times \mathbf{F}_{r_p}$) has approx-dim = geo-dim = p (resp. geo-dim = $p + 1$). Moreover, for every $p \geq 3$, there is a group Γ_p with free actions α_p and β_p such that approx-dim($\mathcal{R}_{\alpha_p}$) = geo-dim($\mathcal{R}_{\alpha_p}$) = p and approx-dim($\mathcal{R}_{\beta_p}$) + 1 = geo-dim($\mathcal{R}_{\beta_p}$) = p.

In [DG09], Dooley-Golodets study the behavior of the dimension geo-dim under finite extensions. The notion of **measurable cohomological dimension** introduced in [ST07] has some similarity with the geometric dimension.

8. L^2-Betti Numbers

The ℓ^2-Betti numbers of cocompact group actions on manifolds were introduced by Atiyah [Ati76] in terms of the heat kernel. Connes [Con79] defined them for measured foliations. Cheeger-Gromov [CG86] introduced ℓ^2-Betti numbers $\beta_n^{(2)}(\Gamma) \in [0, \infty]$, $n \in \mathbb{N}$, for arbitrary countable groups Γ. In [Gab02] the L^2-Betti numbers $\beta_n^{(2)}(\mathcal{R}) \in [0, \infty]$, $n \in \mathbb{N}$ are defined for p.m.p. equivalence relations $\mathcal{R}$, by using proper simplicial $\mathcal{R}$-fields (see section 7). In any case, the definitions rely on the notion of generalized von Neumann dimension, expressed as the trace of certain projections. One of the main results in [Gab00b, Gab02] is the invariance of the $\beta_n^{(2)}(\Gamma)$ under orbit equivalence.

Theorem 8.1 ([Gab02]). *If $\mathcal{R}_\Gamma$ is generated by a free p.m.p. action of Γ, then $\beta_n^{(2)}(\mathcal{R}_\Gamma) = \beta_n^{(2)}(\Gamma)$ for every $n \in \mathbb{N}$.*

The inequality $\mathrm{Cost}(\Gamma) \geq \beta_1^{(2)}(\Gamma) - \beta_0^{(2)}(\Gamma) + 1$ proved in [Gab02] is an equality in all cases where the computations have been achieved, thus leading to the question:

Question 8.2 (Cost vs first ℓ^2-Betti number). Is there an infinite countable group with $\mathrm{Cost}(\Gamma) > \beta_1^{(2)}(\Gamma) + 1$?

The following compression formula was a key point in various places notably when studying "self-similarities" (the "fundamental group", see [Pop06a]) and measure equivalence (see section 9).

Theorem 8.3 ([Gab02]). *The L^2-Betti numbers of $\mathcal{R}$ and of its **restriction** to a Borel subset $A \subset X$ meeting all the classes satisfy: $\beta_n^{(2)}(\mathcal{R}) = \mu(A)\beta_n^{(2)}(\mathcal{R}|A)$.*

It follows that lattices in a common locally compact second countable group have proportional ℓ^2-Betti numbers.

In [BG04], L^2-Betti numbers for profinite actions are used to extend Lück's approximation theorem [Lüc94] to non-normal subgroups. We refer to the book [Lüc02] for information about ℓ^2-Betti numbers of groups and for an alternative approach to von Neumann dimension. See [Sau05, ST07, Tho08] for extension of $\beta_n^{(2)}(\mathcal{R})$ to measured groupoids, and several computations using Lück's approach ([NR09] proves that the various definitions coincide).

Very interesting combinatorial analogues of the cost and $\beta_1^{(2)}$ have been introduced by Elek [Ele07] in a context of sequences of finite graphs.

9. Measure Equivalence

Two groups Γ_1 and Γ_2 are **virtually isomorphic** if there exist $F_i \lhd \Lambda_i < \Gamma_i$ such that $\Lambda_1/F_1 \simeq \Lambda_2/F_2$, where F_i are finite groups, and Λ_i has finite index in Γ_i. This condition is equivalent with: Γ, Λ admit commuting actions on a set Ω such that each of the actions $\Gamma \curvearrowright \Omega$ and $\Lambda \curvearrowright \Omega$ has finite quotient set and finite stabilizers.

A finite set admits two natural generalizations, a topological one (compact set) leading to **geometric group theory** and a measure theoretic one (finite measure set) leading to **measured group theory**.

Definition 9.1 ([Gro93]). Two countable groups Γ_1 and Γ_2 are **measure equivalent (ME)** (**notation:** $\Gamma_1 \overset{\mathrm{ME}}{\sim} \Gamma_2$) if there exist commuting actions of Γ_1 and Γ_2, that are (each) measure preserving, free, and with a finite measure fundamental domain, on some standard (infinite) measure space (Ω, m).

The ratio $[\Gamma_1 : \Gamma_2]_\Omega := m(\Omega/\Gamma_2)/m(\Omega/\Gamma_1)$ of the measures of the fundamental domains is called the **index** of the **coupling** Ω. The typical examples,

besides virtually isomorphic groups, are lattices in a common (locally compact second countable) group G with its Haar measure, acting by left and right multiplication.

The topological analogue was shown to be equivalent with **quasi-isometry (QI)** between finitely generated groups [Gro93], thus raising **measured group theory** (i.e. the study of groups up to ME) to parallel **geometric group theory**. See [Fur99a] for the basis in ME and the surveys [Gab05a, Sha05, Fur09] for more recent developments. Measure equivalence and orbit equivalence are intimately connected by considering the relation between the quotient actions $\Gamma_1 \curvearrowright \Omega/\Gamma_2$ and $\Gamma_2 \curvearrowright \Omega/\Gamma_1$. In fact two groups are ME iff they admit SOE free actions.

Definition 9.2 (Stable Orbit Equivalence). Two p.m.p. actions of $\Gamma_i \curvearrowright (X_i, \mu_i)$ are **stably orbit equivalent (SOE)** if there are Borel subsets $Y_i \subset X_i$, $i = 1, 2$ which meet almost every orbit of Γ_i and a measure-scaling isomorphism $f : Y_1 \to Y_2$ s.t.

$$f(\Gamma_1.x \cap Y_1) = \Gamma_2.f(x) \cap Y_2 \quad \text{a.e.}$$

The **index** or **compression constant** of this SOE f is $[\Gamma_1 : \Gamma_2]_f = \frac{\mu(Y_2)}{\mu(Y_1)}$.

The state of the art ranges from quite well understood ME-classes to mysterious and very rich examples. For instance, the finite groups obviously form a single ME-class. The infinite amenable groups form a single ME-class [OW80]. The ME-class of a lattice in a center-free simple Lie group G with real rank ≥ 2 (like $\mathrm{SL}(n, \mathbb{R})$, $n \geq 3$) consists in those groups that are virtually isomorphic with a lattice in G [Fur99a]. If Γ is a non-exceptional mapping class group, its ME-class consists only in its virtual isomorphism class [Kid08a]. Kida extended this kind of result to some amalgamated free products (see [Kid09]).

On the opposite, the ME-class of the (mutually virtually isomorphic) free groups $\mathbf{F}_r$ $(2 \leq r < \infty)$ contains the free products $*_{i=1}^{r} A_i$ of infinite amenable groups, surface groups $\pi_1(\Sigma_g)$ $(g \geq 2)$, certain branched surface groups [Gab05a], elementarily free groups [BTW07]... and is far from being understood. Being ME with a free group is equivalent to admitting a free p.m.p. treeable action [Hjo06].

There is a considerable list of ME-invariants (see [Gab05a] and the references therein). For instance Kazhdan property (T), Haagerup property, the **ergodic dimension** (resp. **approximate ergodic dimension**) defined as the infimum of the geometric (resp. approximate) dimension among all the free p.m.p. actions of Γ, the sign of the Euler characteristic (when defined), the Cowling-Haagerup invariant, belonging to the classes $\mathcal{C}_{\mathrm{reg}}, \mathcal{C}$. Recently exactness (see [BO08]) and belonging to the class $\mathcal{S}$ of Ozawa [Sak09a] were proved to be ME-invariants. There are also numerical invariants which are preserved under ME modulo multiplication by the index: $\mathrm{Cost}(\Gamma) - 1$, the ℓ^2-Betti numbers $(\beta_n^{(2)}(\Gamma))_{n \in \mathbb{N}}$ [Gab02].

ME is stable under some basic constructions:

(a) if $\Gamma_i \overset{\text{ME}}{\sim} \Lambda_i$ for $i = 1, \cdots, n$ then $\Gamma_1 \times \cdots \times \Gamma_n \overset{\text{ME}}{\sim} \Lambda_1 \times \cdots \times \Lambda_n$

(b) if $\Gamma_i \overset{\text{ME}}{\sim} \Lambda_i$ with index 1, then $\Gamma_1 * \cdots * \Gamma_n \overset{\text{ME}}{\sim} \Lambda_1 * \cdots * \Lambda_n$ (with index 1).

Some papers study when the converse holds [MS06, IPP08, CH10, AG10]. One has of course to impose some irreducibility conditions on the building blocks, and these conditions have to be strong enough to resist the measurable treatment. These requirements are achieved

(a) (for direct products) if the Γ_i, Λ_i belong to the class $\mathcal{C}_{\text{reg}}$ of [MS06] (for instance if they are non-amenable non-trivial free products): the non-triviality of the bounded cohomology $\text{H}^2_{\text{b}}(\Gamma, \ell^2(\Gamma))$ is an ME-invariant preventing Γ to decompose (non-trivially) as a direct product;

(b) (for free products) if the Γ_i, Λ_i are $\mathcal{MFI}$ (for instance if they have $\beta_1^{(2)} = 0$ and are non-amenable) [AG10]: they are not ME with a (non-trivial) free product. We prove for instance:

Theorem 9.3 ([AG10]). *If $\Gamma_1 * \cdots * \Gamma_n \overset{ME}{\sim} \Lambda_1 * \cdots * \Lambda_p$, where both the Γ_i's and the Λ_j's belong to distinct ME-classes and are $\mathcal{MFI}$, then $n = p$ and up to a permutation of the indices $\Gamma_i \overset{ME}{\sim} \Lambda_i$.*

See also [IPP08, CH10] when the groups have Kazhdan property (T), or are direct products, under extra ergodicity hypothesis. The delicate point of removing ergodicity assumptions in [AG10] was achieved by using [Alv09a].

Similar "deconstruction" results were obtained by Sako [Sak09b] for building blocks made of direct products of non-amenable exact groups when considering free products with amalgamation over amenable subgroups or by taking wreath product with amenable base.

Refinements of the notion of ME were introduced in [Sha04, Tho09, LSW09] or by Sauer and Bader-Furman-Sauer.

10. Non-orbit Equivalent Actions for a Given Group

In this section, we only consider ergodic free p.m.p. actions $\Gamma \curvearrowright^{\alpha} (X, \mu)$ of infinite countable groups and the associated orbit equivalence relations $\mathcal{R}_\alpha$. Ornstein-Weiss' theorem [OW80] implies that amenable groups all produce the same relation, namely $\mathcal{R}_{\text{hyp}}$. What about non-amenable groups? How many non-OE actions for a given group? Most of the OE-invariants depend on the group rather than on the action, and thus cannot distinguish between various actions of the group. However, for non-Kazhdan property (T) groups, Connes-Weiss

[CW80] produced two non-OE actions distinguished by strong ergodicity (see section 5). And along the years, various rigidity results entailed some specific families of groups to admit continously[7] many non-OE actions (see for instance [BG81, Zim84, GG88b, MS06, Pop06b, Pop07a]).

We briefly describe below the crucial steps on the route toward the general solution:

Theorem 10.1 ([Ioa07, Eps08]). *Every non-amenable group admits continuously many orbit inequivalent free ergodic p.m.p. actions.*

The first step was made by Hjorth [Hjo05] when, within the circle of ideas from Connes [Con80] and Popa [Pop86], he obtained the result for Kazhdan property (T) groups. Roughly speaking, a pair of OE actions α and β defining the same equivalence relation $\mathcal{R}$ gives a diagonal action $(\gamma.(x,y) = (\gamma._\alpha x, \gamma._\beta y))$ on $\mathcal{R}$ and thus a unitary representation on $L^2(\mathcal{R},\nu)$. When considering uncountably many OE actions, a separability argument shows that the characteristic function $\mathbf{1}_D$ of the diagonal is sufficiently almost invariant for some pair of actions. Now, an invariant vector near $\mathbf{1}_D$, which is given by property (T), delivers a conjugacy between the actions. There exists a continuum of pairwise non-conjugate actions, and by the above the OE-classes in this continuum are countable.

The next step was the analogous theorem for the prototypical non-property (T), non-rigid group, namely the free groups and some free products [GP05]. It lay again within the same circle of ideas but there, rigidity was obtained through Popa's **property (T) relative to the space** (see section 11).

Then Ioana [Ioa07] extended it to all groups containing a copy of $\mathbf{F}_2$. For this, he introduced a weak version of property (T) relative to the space and used a general construction called **co-induction**[8].

Eventually, Epstein obtained the theorem in full generality [Eps08]. For this she had to generalize the co-induction construction to the setting provided by Gaboriau-Lyons' measurable solution to **von Neumann's problem** (see below). Moreover, Ioana extended Epstein's result from orbit inequivalent to von Neumann inequivalent actions [Ioa07].

When von Neumann introduced the notion of amenability [vN29], he observed that a countable group containing a copy of $\mathbf{F}_2$ cannot be amenable. The question of knowing whether every non-amenable countable group has to contain a copy of $\mathbf{F}_2$, known as **von Neumann's problem**, was answered in the negative by Ol′šanskiĭ [Ol′80]. In the measurable framework, offering much more flexibility, the answer is somewhat different:

[7] This is an upper bound since $\mathrm{Card}(\mathrm{Aut}([0,1],\mathrm{Leb})) = 2^{\aleph_0}$.

[8] Co-induction is the classical right adjoint of restriction. Its measure theoretic version was brought to my attention by Sauer and used in [Gab05a], but it probably first appeared in preliminary versions of [DGRS08].

Theorem 10.2 ([GL09]). *For any non-amenable countable group Γ, the orbit equivalence relation of the Bernoulli shift action $\Gamma \curvearrowright ([0,1],\mathrm{Leb})^\Gamma$ contains a subrelation generated by a free ergodic p.m.p. action of $\mathbf{F}_2$.*

In the terminology of [Mon06], there is a randembedding of $\mathbf{F}_2$ in any non-amenable group. The proof uses percolation theory on graphs and [HP99, LS99, PSN00, Gab05b, Hjo06]. The following general question remains open:

Question 10.3. Does every ergodic non-hyperfinite p.m.p. equivalence relation contain a (treeable) subrelation of cost > 1?

11. Relative Property (T)

In his seminal paper [Kaz67] on property (T), Kazhdan implicitly[9] introduced the notion of property (T) relative to a subgroup $\Lambda < \Gamma$. In particular, a group always has property (T) relative to its "unit subgroup" $\{1\} < \Gamma$. When considering a groupoid like $\mathcal{R}$, its space of units (X,μ) (and its "relative representation theory") is much more complicated. The introduction by Popa [Pop06a] of the fruitful notion of **property (T) relative to the space** (X,μ) (also simply called **rigidity**) allowed him to solve some long standing problems in von Neumann algebras. In fact, the definition involves a pair of von Neumann algebras $B \subset M$ (for instance $L^\infty(X,\mu) \subset L(\mathcal{R})$) and parallels the analogous notion for groups, in the spirit of Connes-Jones [CJ85].

The typical example is provided by the standard action of $\mathrm{SL}(2,\mathbb{Z})$ and its non-amenable subgroups Γ (for instance free groups $\mathbf{F}_r$, $r \geq 2$) on $\mathbb{T}^2$. Notice Ioana's result that in fact every ergodic non-amenable subrelation of $\mathcal{R}_{\mathrm{SL}(2,\mathbb{Z})\curvearrowright\mathbb{T}^2}$ still has property (T) relative to the space $\mathbb{T}^2$ [Ioa09]. The property (T) relative to the space (X,μ) comes from the group property (T) of $\mathbb{Z}^2 \rtimes \Gamma$ relative to the subgroup $\mathbb{Z}^2$, via viewing $\mathbb{Z}^2$ as the Pontryagin dual of $\mathbb{T}^2$. This property (never satisfied by standard Bernoulli shifts) entails several rigidity phenomena (see for instance [Pop06a, IPP08, GP05]). More examples come from [Val05, Fer06] and they all involve some arithmeticity. This led Popa to ask for the class of groups admitting such a free p.m.p. action with property (T) relative to the space. Törnquist [Tör06] ensures that the class is stable under taking a free product with any countable group. More generally, [Gab08] shows that the class contains all the non-trivial free products of groups $\Gamma = \Gamma_1 * \Gamma_2$: in fact $\mathcal{R}_{\Gamma_1}$ and $\mathcal{R}_{\Gamma_2}$ may be chosen to be conjugate with any prescribed free Γ_i-action and the arithmeticity alluded to is hidden in the way they are put in free product. This leads, using ideas from [PV08a] to (plenty of) examples of $\mathcal{R}_\Gamma$ with trivial outer automorphism group, in particular the first examples for free $\mathbf{F}_2$-actions [Gab08]. Ioana [Ioa07] proved that every non-amenable group admits a free p.m.p. action satisfying a weak form of the above property, enough for various purposes, see section 10.

[9] This was made explicit in [Mar82].

12. Some Rigidity Results

We have three notions of equivalence between free p.m.p. actions:

$$(\Gamma_1 \curvearrowright^{\alpha_1} X_1 \overset{\text{Conj}}{\sim} \Gamma_2 \curvearrowright^{\alpha_2} X_2) \implies (\Gamma_1 \curvearrowright^{\alpha_1} X_1 \overset{\text{OE}}{\sim} \Gamma_2 \curvearrowright^{\alpha_2} X_2) \implies (\mathcal{R}_{\alpha_1} \overset{\text{vN}}{\sim} \mathcal{R}_{\alpha_2}).$$

Rigidity phenomena consist ideally in situations where (for free actions) some implication can be reversed, or more generally when a big piece of information of a stronger nature can be transferred through a weaker equivalence. Zimmer's pioneering work (see [Zim84]) inaugurated a series of impressive results of rigidity for the first arrow ($\overset{\text{Conj}}{\sim} \dashleftarrow \overset{\text{OE}}{\sim}$), made possible by the introduction in OE theory and in operator algebras of new techniques borrowed from diverse mathematical domains, like algebraic groups, geometry, geometric group theory, representation theory or operator algebras. These rigidity results for $\Gamma_1 \curvearrowright^{\alpha_1} X_1$ take various qualifications according to whether an OE hypothesis entails

- **strong OE rigidity**: conjugacy under some additionnal hypothesis about the mysterious action $\Gamma_2 \curvearrowright^{\alpha_2} X_2$, or even
- **OE superrigidity**: conjugacy of the actions with no hypothesis at all on the target action.

These notions are **virtual** when they happen only up to finite groups (see [Fur99b] for precise definitions).

To give some ideas we simply evoke a sample of some typical and strong statements far from exhaustiveness or full generality.

Theorem 12.1 ([Fur99b]). *Any free action that is OE with the standard action* $\mathrm{SL}(n,\mathbb{Z}) \curvearrowright \mathbb{T}^n$ *for* $n \geq 3$*, is virtually conjugate with it.*

This is more generally true for lattices in a connected, center-free, simple, Lie group of higher rank, and for "generic" actions (see [Fur99b]). Monod-Shalom [MS06] obtained strong OE rigidity results when Γ_1 is a direct product of groups in $\mathcal{C}_{\text{reg}}$, under appropriate ergodicity assumptions on both sides. See also Hjorth-Kechris [HK05] for rigidity results about actions of products, where the focus is more on Borel reducibility. Kida's results [Kid08b] consider actions of mapping class groups of orientable surfaces and their direct products. He also obtains very strong rigidity results for certain amalgamated free products [Kid09]. A series of ground breaking results in von Neumann algebras obtained by Popa [Pop06a, Pop06c, Pop06d, Pop07a, Pop08] and his collaborators [PS07, IPP08, PV08d, PV08b, PV08a, Ioa08, PV08c, PV09b] (see [Vae07] for a review) dramatically modified the landscape. On the OE side, these culminated in Popa's cocycle superrigidity theorems, that imply several impressive OE superrigidity corollaries, for instance:

Theorem 12.2 ([Pop07a, Pop08]). *Assume that Γ is either an infinite ICC Kazhdan property (T) group or is the product of two infinite groups $H \times H'$ and has no finite normal subgroup. Then any free action that is orbit equivalent with the Bernoulli shift $\Gamma \curvearrowright (X_0, \mu_0)^\Gamma$ is conjugate with it.*

See Furman's ergodic theoretical treatment and generalizations [Fur07] for the Kazhdan property (T) case. In the opposite direction, Bowen obtained some surprising non-rigidity results [Bow09a, Bow09b] showing for instance that all the Bernoulli shifts of the free groups $\mathbf{F}_r$, $2 \leq r < \infty$ are mutually SOE (see Def. 9.2).

As it follows from [Sin55, FM77b], being able to reverse the second arrow ($\overset{\text{OE}}{\sim} \dashleftarrow \overset{\text{vN}}{\sim}$) essentially amounts to being able to uniquely identify the Cartan subalgebra inside $L(\mathcal{R})$, i.e. given two Cartan subalgebras A_1, A_2 in $L(\mathcal{R}_1) \simeq L(\mathcal{R}_2)$, being able to relate them through the isomorphism. Such results are qualified **vNE rigidity** or **W^*-rigidity**. The starting point is Popa's breakthrough [Pop06a] where a uniqueness result is obtained under some hypothesis on both A_1 and A_2 (and this was enough to solve long standing problems in von Neumann algebras). See also [IPP08, CH10] for this kind of strong statements under various quite general conditions. We refer to the surveys [Pop07b, Vae07, Vae10] for the recent developments in vNE or W^*-rigidity. However, after a series of progresses (see for instance [OP08a, OP08b, Ioa08, Pet09, PV09b, Pet10]), the most recent achievement is:

Theorem 12.3 ([Ioa10]). *If a free action of a group is von Neumann equivalent with the standard Bernoulli shift action of an ICC Kazhdan property (T) group, then the actions are in fact conjugate.*

13. Some Further OE-invariants

In order to distinguish treeable Borel equivalence relations, Hjorth introduced a technique preventing a p.m.p. equivalence relation from being OE with a profinite one [Hjo06]. Then Kechris and Epstein-Tsankov isolated representation-theoretic properties (i.e. in terms of the Koopman representation) leading to strong forms of non-profiniteness; see [Kec05, ET10].

Elek-Lippner introduced the **sofic** property for equivalence relations. It is satisfied by profinite actions, treeable equivalence relations and Bernoulli shifts of sofic groups [EL10]. They also proved that the associated von Neumann algebra satisfies the Connes' embedding conjecture.

Acknowledgements

I'm grateful to A. Alvarez, C. Houdayer and J. Melleray for their comments.

References

[Ada88] S. Adams. Indecomposability of treed equivalence relations. *Israel J. Math.*, 64(3):362–380 (1989), 1988.

[Ada90] S. Adams. Trees and amenable equivalence relations. *Ergodic Theory Dynamical Systems*, 10(1):1–14, 1990.

[ADR00] C. Anantharaman-Delaroche and J. Renault. *Amenable groupoids*, volume 36 of *Monographies de L'Enseignement Mathématique*. L'Enseignement Mathématique, Geneva, 2000.

[AG10] A. Alvarez and D. Gaboriau. Free products, orbit equivalence and measure equivalence rigidity. *to appear in G.G.D.*, 2010.

[Alv09a] A. Alvarez. Un théorème de Kurosh pour les relations d'équivalence boréliennes. *to appear in Annales de l'Institut Fourier*, 2009.

[Alv09b] A. Alvarez. Une thrie de Bass-Serre pour les groupoes boriens. *preprint*, 2009.

[AN07] M. Abert and N. Nikolov. Rank gradient, cost of groups and the rank versus Heegaard genus problem. *preprint*, 2007.

[AS90] S. Adams and R. Spatzier. Kazhdan groups, cocycles and trees. *Amer. J. Math.*, 112(2):271–287, 1990.

[Ati76] M. Atiyah. Elliptic operators, discrete groups and von Neumann algebras. In *Colloque "Analyse et Topologie" en l'Honneur de Henri Cartan (Orsay, 1974)*, pages 43–72. Astérisque, SMF, No. 32–33. Soc. Math. France, Paris, 1976.

[AW] M. Abert and B. Weiss. Bernoulli actions are weakly contained in any free action. in preparation.

[Bas76] H. Bass. Some remarks on group actions on trees. *Comm. Algebra*, 4(12):1091–1126, 1976.

[BG80] S. I. Bezuglyĭ and V. Ja. Golodec. Topological properties of complete groups of automorphisms of a space with a measure. *Siberian Math. J.*, 21(2):147–155, 1980.

[BG81] S. I. Bezuglyĭ and V. Ya. Golodets. Hyperfinite and II_1 actions for nonamenable groups. *J. Funct. Anal.*, 40(1):30–44, 1981.

[BG04] N. Bergeron and D. Gaboriau. Asymptotique des nombres de Betti, invariants l^2 et laminations. *Comment. Math. Helv.*, 79(2):362–395, 2004.

[BO08] N. P. Brown and N. Ozawa. *C^*-algebras and finite-dimensional approximations*, volume 88 of *Graduate Studies in Mathematics*. A.M.S., 2008.

[Bow09a] L. Bowen. Orbit equivalence, coinduced actions and free products. *to appear in Groups Geom. Dyn.*, 2009.

[Bow09b] L. Bowen. Stable orbit equivalence of Bernoulli shifts over free groups. *to appear in Groups Geom. Dyn.*, 2009.

[Bro82] K. Brown. *Cohomology of groups*. Springer-Verlag, New York, 1982.

[BTW07] M. R. Bridson, M. Tweedale, and H. Wilton. Limit groups, positive-genus towers and measure-equivalence. *Ergodic Theory Dynam. Systems*, 27(3):703–712, 2007.

[CFW81] A. Connes, J. Feldman, and B. Weiss. An amenable equivalence relation is generated by a single transformation. *Ergodic Theory Dynamical Systems*, 1(4):431–450 (1982), 1981.

[CG86] J. Cheeger and M. Gromov. L_2-cohomology and group cohomology. *Topology*, 25(2):189–215, 1986.

[CH10] I. Chifan and C. Houdayer. Bass-Serre rigidity results in von Neumann algebras. *to appear in Duke Math. J.*, 2010.

[CI10] I. Chifan and A. Ioana. Ergodic subequivalence relations induced by a Bernoulli action. *to appear in Geom. and Funct. Anal.*, 2010.

[CJ82] A. Connes and V. Jones. A II_1 factor with two nonconjugate Cartan subalgebras. *Bull. Amer. Math. Soc. (N.S.)*, 6(2):211–212, 1982.

[CJ85] A. Connes and V. Jones. Property T for von Neumann algebras. *Bull. London Math. Soc.*, 17(1):57–62, 1985.

[Con79] A. Connes. Sur la théorie non commutative de l'intégration. In *Algèbres d'opérateurs (Sém., Les Plans-sur-Bex, 1978)*, pages 19–143. Springer, Berlin, 1979.

[Con80] A. Connes. A factor of type II_1 with countable fundamental group. *J. Operator Theory*, 4(1):151–153, 1980.

[Con04] A. Connes. Nombres de Betti L^2 et facteurs de type II_1 (d'après D. Gaboriau et S. Popa). Number 294, pages ix, 321–333. S.M.F., 2004.

[CW80] A. Connes and B. Weiss. Property T and asymptotically invariant sequences. *Israel J. Math.*, 37(3):209–210, 1980.

[DG09] A. H. Dooley and V. Ya. Golodets. The geometric dimension of an equivalence relation and finite extensions of countable groups. *Ergodic Theory Dynam. Systems*, 29(6):1789–1814, 2009.

[DGRS08] A. H. Dooley, V. Ya. Golodets, D. J. Rudolph, and S. D. Sinel′shchikov. Non-Bernoulli systems with completely positive entropy. *Ergodic Theory Dynam. Systems*, 28(1):87–124, 2008.

[Dye59] H. Dye. On groups of measure preserving transformation. I. *Amer. J. Math.*, 81:119–159, 1959.

[Dye63] H. Dye. On groups of measure preserving transformations. II. *Amer. J. Math.*, 85:551–576, 1963.

[EL10] G. Elek and G. Lippner. Sofic equivalence relations. *J. Funct. Anal.*, 258(5):1692–1708, 2010.

[Ele07] G. Elek. The combinatorial cost. *Enseign. Math. (2)*, 53(3-4):225–235, 2007.

[EM09] I. Epstein and N. Monod. Nonunitarizable representations and random forests. *Int. Math. Res. Not. IMRN*, (22):4336–4353, 2009.

[Eps08] I. Epstein. Orbit inequivalent actions of non-amenable groups. *preprint*, 2008.

[ET10] I. Epstein and T. Tsankov. Modular actions and amenable representations. *Trans. Amer. Math. Soc.*, 362(2):603–621, 2010.

[Fer06] T. Fernós. Relative property (T) and linear groups. *Ann. Inst. Fourier (Grenoble)*, 56(6):1767–1804, 2006.

[FM77a] J. Feldman and C. Moore. Ergodic equivalence relations, cohomology, and von Neumann algebras. I. *Trans. Amer. Math. Soc.*, 234(2):289–324, 1977.

[FM77b] J. Feldman and C. Moore. Ergodic equivalence relations, cohomology, and von Neumann algebras. II. *Trans. Amer. Math. Soc.*, 234(2):325–359, 1977.

[Fur99a] A. Furman. Gromov's measure equivalence and rigidity of higher rank lattices. *Ann. of Math. (2)*, 150(3):1059–1081, 1999.

[Fur99b] A. Furman. Orbit equivalence rigidity. *Ann. of Math. (2)*, 150(3):1083–1108, 1999.

[Fur05] A. Furman. Outer automorphism groups of some ergodic equivalence relations. *Comment. Math. Helv.*, 80(1):157–196, 2005.

[Fur07] A. Furman. On Popa's cocycle superrigidity theorem. *Int. Math. Res. Not. IMRN*, (19):Art. ID rnm073, 46, 2007.

[Fur09] A. Furman. A survey of measured group theory. *Proceedings of a Conference honoring Robert Zimmer's 60th birthday, arXiv:0901.0678v1 [math.DS]*, 2009.

[Gab98] D. Gaboriau. Mercuriale de groupes et de relations. *C. R. Acad. Sci. Paris Sér. I Math.*, 326(2):219–222, 1998.

[Gab00a] D. Gaboriau. Coût des relations d'équivalence et des groupes. *Invent. Math.*, 139(1):41–98, 2000.

[Gab00b] D. Gaboriau. Sur la (co-)homologie L^2 des actions préservant une mesure. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(5):365–370, 2000.

[Gab02] D. Gaboriau. Invariants L^2 de relations d'équivalence et de groupes. *Publ. Math. Inst. Hautes Études Sci.*, 95:93–150, 2002.

[Gab05a] D. Gaboriau. Examples of groups that are measure equivalent to the free group. *Ergodic Theory Dynam. Systems*, 25(6):1809–1827, 2005.

[Gab05b] D. Gaboriau. Invariant percolation and harmonic Dirichlet functions. *Geom. Funct. Anal.*, 15(5):1004–1051, 2005.

[Gab08] D. Gaboriau. Relative property (T) actions and trivial outer automorphism groups. *preprint*, 2008.

[Gab10a] D. Gaboriau. Approximations of equivalence relations. *in preparation*, 2010.

[Gab10b] D. Gaboriau. What is cost. *preprint*, 2010.

[Gef93] S. L. Gefter. Ergodic equivalence relation without outer automorphisms. *Dopov./Dokl. Akad. Nauk Ukraïni*, 11:25–27, 1993.

[Gef96] S. L. Gefter. Outer automorphism group of the ergodic equivalence relation generated by translations of dense subgroup of compact group on its homogeneous space. *Publ. Res. Inst. Math. Sci.*, 32(3):517–538, 1996.

[GG88a] S. L. Gefter and V. Ya. Golodets. Fundamental groups for ergodic actions and actions with unit fundamental groups. *Publ. Res. Inst. Math. Sci.*, 24(6):821–847 (1989), 1988.

[GG88b] S. L. Gefter and V. Ya. Golodets. Fundamental groups for ergodic actions and actions with unit fundamental groups. *Publ. Res. Inst. Math. Sci.*, 24(6):821–847 (1989), 1988.

[Ghy95] É. Ghys. Topologie des feuilles génériques. *Ann. of Math. (2)*, 141(2):387–422, 1995.

[GL09] D. Gaboriau and R. Lyons. A measurable-group-theoretic solution to von Neumann's problem. *Invent. Math.*, 177(3):533–540, 2009.

[GP05] D. Gaboriau and S. Popa. An uncountable family of nonorbit equivalent actions of $\mathbf{F}_n$. *J. Amer. Math. Soc.*, 18(3):547–559 (electronic), 2005.

[GP07] T. Giordano and V. Pestov. Some extremely amenable groups related to operator algebras and ergodic theory. *J. Inst. Math. Jussieu*, 6(2):279–315, 2007.

[Gro93] M. Gromov. Asymptotic invariants of infinite groups. In *Geometric group theory, Vol. 2 (Sussex, 1991)*, volume 182 of *London Math. Soc. Lecture Note Ser.*, pages 1–295. Cambridge Univ. Press, Cambridge, 1993.

[Gro00] M. Gromov. Spaces and questions. *Geom. Funct. Anal.*, (Special Volume, Part I):118–161, 2000. GAFA 2000 (Tel Aviv, 1999).

[Hjo05] G. Hjorth. A converse to Dye's theorem. *Trans. Amer. Math. Soc.*, 357(8):3083–3103 (electronic), 2005.

[Hjo06] G. Hjorth. A lemma for cost attained. *Ann. Pure Appl. Logic*, 143(1-3):87–102, 2006.

[HK05] G. Hjorth and A. S. Kechris. Rigidity theorems for actions of product groups and countable Borel equivalence relations. *Mem. Amer. Math. Soc.*, 177(833):viii+109, 2005.

[HP99] O. Häggström and Y. Peres. Monotonicity of uniqueness for percolation on Cayley graphs: all infinite clusters are born simultaneously. *Probab. Theory Related Fields*, 113(2):273–285, 1999.

[Ioa07] A. Ioana. Orbit inequivalent actions for groups containing a copy of $\mathbf{F}_2$. *preprint*, 2007.

[Ioa08] A. Ioana. Cocycle superrigidity for profinite action of property (T) groups. *Prublication*, 2008.

[Ioa09] A. Ioana. Relative property (T) for the subequivalence relations induced by the action of $\mathrm{SL}_2(\mathbf{Z})$ on $\mathbf{T^2}$. *preprint*, 2009.

[Ioa10] A. Ioana. W*-superrigidity for Bernoulli actions of property (T) groups, 2010.

[IPP08] A. Ioana, J. Peterson, and S. Popa. Amalgamated free products of weakly rigid factors and calculation of their symmetry groups. *Acta Math.*, 200(1):85–153, 2008.

[JKL02] S. Jackson, A. S. Kechris, and A. Louveau. Countable Borel equivalence relations. *J. Math. Log.*, 2(1):1–80, 2002.

[JS87] Vaughan F. R. Jones and Klaus Schmidt. Asymptotically invariant sequences and approximate finiteness. *Amer. J. Math.*, 109(1):91–114, 1987.

[Kaz67] D. A. Kazhdan. On the connection of the dual space of a group with the structure of its closed subgroups. *Funkcional. Anal. i Priložen.*, 1:71–74, 1967.

[Kec05] A. S. Kechris. Unitary representations and modular actions. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 326(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 13):97–144, 281–282, 2005.

[Kec10] A. S. Kechris. *Global aspects of ergodic group actions*, volume 160 of *Mathematical Surveys and Monographs*. A.M.S., 2010.

[Kid08a] Y. Kida. The mapping class group from the viewpoint of measure equivalence theory. *Mem. Amer. Math. Soc.*, 196(916):viii+190, 2008.

[Kid08b] Y. Kida. Orbit equivalence rigidity for ergodic actions of the mapping class group. *Geom. Dedicata*, 131:99–109, 2008.

[Kid08c] Y. Kida. Outer automorphism groups of equivalence relations for mapping class group actions. *J. Lond. Math. Soc. (2)*, 78(3):622–638, 2008.

[Kid09] Y. Kida. Rigidity in measure-theoretic group theory for amalgamated free products. *Preprint*, 2009.

[KM04] A. S. Kechris and B. D. Miller. *Topics in orbit equivalence*, volume 1852 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.

[Koo31] B. O. Koopman. Hamiltonian systems and transformations in Hilbert space. *Proceedings USA Academy*, 17:315–318, 1931.

[KT08] A. S. Kechris and T. Tsankov. Amenable actions and almost invariant sets. *Proc. Amer. Math. Soc.*, 136(2):687–697 (electronic), 2008.

[KT10] J. Kittrell and T. Tsankov. Topological properties of full groups. *Ergodic Theory Dynam. Systems, to appear*, 2010.

[Lev95] G. Levitt. On the cost of generating an equivalence relation. *Ergodic Theory Dynam. Systems*, 15(6):1173–1181, 1995.

[LP09] R. Lyons and Y. Peres. *Probability on Trees and Networks*. Cambridge University Press, In preparation, Cambridge, 2009.

[LPV08] R. Lyons, M. Pichot, and S. Vassout. Uniform non-amenability, cost, and the first l^2-Betti number. *Groups Geom. Dyn.*, 2(4):595–617, 2008.

[LS99] R. Lyons and O. Schramm. Indistinguishability of percolation clusters. *Ann. Probab.*, 27(4):1809–1836, 1999.

[LSW09] W. Lück, R. Sauer, and C. Wegner. L2-torsion, the measure-theoretic determinant conjecture, and uniform measure equivalence. *preprint*, 2009.

[Lüc94] W. Lück. Approximating L^2-invariants by their finite-dimensional analogues. *Geom. Funct. Anal.*, 4(4):455–481, 1994.

[Lüc02] W. Lück. *L^2-invariants: theory and applications to geometry and K-theory*, volume 44. Springer-Verlag, Berlin, 2002.

[Mar82] G. A. Margulis. Finitely-additive invariant measures on Euclidean spaces. *Ergodic Theory Dynam. Systems*, 2(3–4):383–396 (1983), 1982.

[MO10] N. Monod and N. Ozawa. The Dixmier problem, lamplighters and Burnside groups. *J. Funct. Anal.*, 258(1):255–259, 2010.

[Mon06] N. Monod. An invitation to bounded cohomology. In *International Congress of Mathematicians. Vol. II*, pages 1183–1211. Eur. Math. Soc., Zürich, 2006.

[Moo82] C. C. Moore. Ergodic theory and von Neumann algebras. In *Operator algebras and applications, Part 2 (Kingston, Ont., 1980)*, pages 179–226. Amer. Math. Soc., Providence, R.I., 1982.

[MS06] N. Monod and Y. Shalom. Orbit equivalence rigidity and bounded cohomology. *Ann. of Math. (2)*, 164(3):825–878, 2006.

[MvN36] F. Murray and J. von Neumann. On rings of operators. *Ann. of Math., II. Ser.*, 37:116–229, 1936.

[NR09] S. Neshveyev and S. Rustad. On the definition of L^2-Betti numbers of equivalence relations. *Internat. J. Algebra Comput.*, 19(3):383–396, 2009.

[Ol′80] A. Ju. Ol′šanskiĭ. On the question of the existence of an invariant mean on a group. *Uspekhi Mat. Nauk*, 35(4(214)):199–200, 1980.

[OP08a] N. Ozawa and S. Popa. On a class of II_1 factors with at most one cartan subalgebra. *Ann. of Math., to appear*, 2008.

[OP08b] N. Ozawa and S. Popa. On a class of II_1 factors with at most one cartan subalgebra II, 2008.

[OW80] D. Ornstein and B. Weiss. Ergodic theory of amenable group actions. I. The Rohlin lemma. *Bull. Amer. Math. Soc. (N.S.)*, 2(1):161–164, 1980.

[Oza04] N. Ozawa. Solid von Neumann algebras. *Acta Math.*, 192(1):111–117, 2004.

[Oza06] N. Ozawa. A Kurosh-type theorem for type II_1 factors. *Int. Math. Res. Not.*, pages Art. ID 97560, 21, 2006.

[Oza09] N. Ozawa. An example of a solid von Neumann algebra. *Hokkaido Math. J.*, 38(3):557–561, 2009.

[Pau99] F. Paulin. Propriétés asymptotiques des relations d'équivalences mesurées discrètes. *Markov Process. Related Fields*, 5(2):163–200, 1999.

[Pet09] J. Peterson. L^2-rigidity in von Neumann algebras. *Invent. Math.*, 175(2):417–433, 2009.

[Pet10] J. Peterson. Examples of group actions which are virtually w*-superrigid, 2010.

[Pic07a] M. Pichot. Espaces mesurés singuliers fortement ergodiques (Étude métrique-mesurée). *Ann. Inst. Fourier (Grenoble)*, 57(1):1–43, 2007.

[Pic07b] M. Pichot. Sur la théorie spectrale des relations d'équivalence mesurées. *J. Inst. Math. Jussieu*, 6(3):453–500, 2007.

[Pop86] S. Popa. Correspondences. *INCREST Preprint, 56*, 1986.

[Pop06a] S. Popa. On a class of type II_1 factors with Betti numbers invariants. *Ann. of Math. (2)*, 163(3):809–899, 2006.

[Pop06b] S. Popa. Some computations of 1-cohomology groups and construction of non-orbit-equivalent actions. *J. Inst. Math. Jussieu*, 5(2):309–332, 2006.

[Pop06c] S. Popa. Strong rigidity of II_1 factors arising from malleable actions of w-rigid groups. I. *Invent. Math.*, 165(2):369–408, 2006.

[Pop06d] S. Popa. Strong rigidity of II_1 factors arising from malleable actions of w-rigid groups. II. *Invent. Math.*, 165(2):409–451, 2006.

[Pop07a] S. Popa. Cocycle and orbit equivalence superrigidity for malleable actions of w-rigid groups. *Invent. Math.*, 170(2):243–295, 2007.

[Pop07b] S. Popa. Deformation and rigidity for group actions and von Neumann algebras. In *I.C.M. Vol. I*, pages 445–477. Eur. Math. Soc., Zürich, 2007.

[Pop08] S. Popa. On the superrigidity of malleable actions with spectral gap. *J. Amer. Math. Soc.*, 21(4):981–1000, 2008.

[PS07] S. Popa and R. Sasyk. On the cohomology of Bernoulli actions. *Ergodic Theory Dynam. Systems*, 27(1):241–251, 2007.

[PSN00] I. Pak and T. Smirnova-Nagnibeda. On non-uniqueness of percolation on nonamenable Cayley graphs. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(6):495–500, 2000.

[PV08a] S. Popa and S. Vaes. Actions of $\mathbf{F}_\infty$ whose II_1 factors and orbit equivalence relations have prescribed fundamental group. *preprint*, 2008.

[PV08b] S. Popa and S. Vaes. Cocycle and orbit superrigidity for lattices in $\mathrm{SL}(n, \mathbb{R})$ acting on homogeneous spaces, 2008.

[PV08c] S. Popa and S. Vaes. On the fundamental group of II_1 factors and equivalence relations arising from group actions, 2008.

[PV08d] S. Popa and S. Vaes. Strong rigidity of generalized Bernoulli actions and computations of their symmetry groups. *Adv. Math.*, 217(2):833–872, 2008.

[PV09a] M. Pichot and S. Vassout. Le coût est un invariant isopérimétrique. *to appear in Journal of Non comm. Geom.*, 2009.

[PV09b] S. Popa and S. Vaes. Group measure space decomposition of II_1 factors and W*-superrigidity, 2009.

[Sak09a] H. Sako. The class S as an ME invariant. *Int. Math. Res. Not. IMRN*, (15):2749–2759, 2009.

[Sak09b] H. Sako. Measure equivalence rigidity and bi-exactness of groups. *J. Funct. Anal.*, 257(10):3167–3202, 2009.

[Sau05] R. Sauer. L^2-Betti numbers of discrete measured groupoids. *Internat. J. Algebra Comput.*, 15(5-6):1169–1188, 2005.

[Sau09] R. Sauer. Amenable covers, volume and L^2-Betti numbers of aspherical manifolds. *J. Reine Angew. Math.*, 636:47–92, 2009.

[Sch80] K. Schmidt. Asymptotically invariant sequences and an action of $\mathrm{SL}(2, \mathbf{Z})$ on the 2-sphere. *Israel J. Math.*, 37(3):193–208, 1980.

[Sch81] K. Schmidt. Amenability, Kazhdan's property T, strong ergodicity and invariant means for ergodic group-actions. *Ergodic Theory Dynam. Systems*, 1(2):223–236, 1981.

[Ser77] J.-P. Serre. *Arbres, amalgames, SL_2*, volume 46 of *Astérisque*. S.M.F., 1977.

[Sha04] Y. Shalom. Harmonic analysis, cohomology, and the large-scale geometry of amenable groups. *Acta Math.*, 192(2):119–185, 2004.

[Sha05] Y. Shalom. Measurable group theory. In *European Congress of Mathematics*, pages 391–423. Eur. Math. Soc., Zürich, 2005.

[Sin55] I. M. Singer. Automorphisms of finite factors. *Amer. J. Math.*, 77:117–133, 1955.

[ST07] R. Sauer and A. Thom. A spectral sequence to compute L^2-betti numbers of groups and groupoids. 2007.

[Tho08] A. Thom. L^2-invariants and rank metric. In *C^*-algebras and elliptic theory II*, Trends Math., pages 267–280. Birkhäuser, Basel, 2008.

[Tho09] A. Thom. Low degree bounded cohomology and L^2-invariants for negatively curved groups. *Groups Geom. Dyn.*, 3(2):343–358, 2009.

[Tör06] A. Törnquist. Orbit equivalence and actions of $\mathbf{F_n}$. *J. Symbolic Logic*, 71(1):265–282, 2006.

[Vae07] S. Vaes. Rigidity results for Bernoulli actions and their von Neumann algebras (after Sorin Popa). Number 311, pages Exp. No. 961, viii, 237–294. S.M.F., 2007. Séminaire Bourbaki. Vol. 2005/2006.

[Vae10] S. Vaes. Rigidity for von neumann algebras and their invariants. *Proceedings of the International Congress of Mathematicians, Hyderabad, India*, 2010.

[Val05] A. Valette. Group pairs with property (T), from arithmetic lattices. *Geom. Dedicata*, 112:183–196, 2005.

[vN29] J. von Neumann. Zur allgemeinen theorie des maßes. *Fund. Math.*, 13:73–116, 1929.

[Zim84] R. J. Zimmer. *Ergodic theory and semisimple groups*, volume 81 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 1984.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Group Actions on Operator Algebras

Masaki Izumi*

Abstract

We give a brief account of group actions on operator algebras mainly focusing on classification results. We first recall rather classical results on the classification of discrete amenable group actions on the injective factors, which may serve as potential goals in the case of C^*-algebras for the future. We also mention Galois correspondence type results and quantum group actions for von Neumann algebras. Then we report on the recent developments of the classification of group actions on C^*-algebras in terms of K-theoretical invariants. We give conjectures on the classification of a class of countable amenable group actions on Kirchberg algebras and strongly self-absorbing C^*-algebras, which involve the classifying spaces of the groups.

Mathematics Subject Classification (2010). Primary 46L40; Secondary 46L35.

Keywords. Operator algebras, group actions, K-theory

1. Introduction

There are two classes of main objects in the theory of operator algebras, C^*-algebras and von Neumann algebras. They are subalgebras of the set of bounded operators $B(H)$ on a complex Hilbert space H closed under the adjoint operation, and closed under appropriate topologies, the norm topology for the former, and the weak operator topology for the latter. Since the celebrated Gelfand-Naimark theorem says that any abelian C^*-algebra with unit is isomorphic to the set of continuous functions $C(X)$ on a compact Hausdorff space X, C^*-algebras are sometimes regarded as noncommutative analogues of topological spaces, while von Neumann algebras are regarded as those of measure spaces for a similar reason. To certain extent, this analogy is helpful to understand the difference between the two classes, though it may be misleading sometimes. If one further pursues the analogy, the difference between group actions on

*Graduate School of Science, Kyoto University, Sakyo-ku, Kyoto, 606-8502, Japan.
E-mail: izumi@math.kyoto-u.ac.jp

C^*-algebras and those on von Neumann algebras could be compared to that between topological dynamics and ergodic theory. Group actions on operator algebras have always been one of the main interests in the field for the sake of applications to physics, and of course, for intrinsic reasons.

In operator algebraic formulation of quantum physics, the symmetries and time evolution of a physical system are usually described by (anti-)automorphisms of a relevant operator algebra. Indeed, for the purpose of direct applications to physics, a group of mathematical physicists started working on group actions on operator algebras in the 60s, which brought important new ideas to the field such as the KMS-condition for actions of the real numbers $\mathbb{R}$ (see [1],[3],[11] for example).

Group actions are also essential for understanding the structure of operator algebras. One can see it typically in Connes's classification of injective factors, which involves the classification of cyclic group actions. Factors are von Neumann algebras with trivial center, and they are building blocks of general von Neumann algebras. Connes completely classified injective factors up to isomorphism in the mid 70s except for one case, which was settled by Haagerup about 10 years later. Thanks to these results, it turns out that injectivity, which is a functional analytic property, is equivalent to approximately finite dimensionality (see [52]).

Connes's argument for the classification of cyclic group actions is based on a noncommutative analogue of the Rohlin tower construction in ergodic theory, which has a great influence on other classification results of group actions. A far reaching generalization of Connes's classification of cyclic group actions was accomplished by many hands, which says that countable amenable group actions on injective factors are completely classified up to cocycle conjugacy by computable classification invariants (see [23]).

These results on injective factors and countable amenable group actions on them are one of the most significant establishments in the theory of operator algebras, which may suggest possible goals in other areas of operator algebras for the future. One of the purposes of this note is to report on the progress of group actions on operator algebras after these results. We mainly focus on the case of C^*-algebras, though we also mention other topics such as Galois correspondence for compact group actions and quantum group actions in the von Neumann algebra case.

Nuclearity for C^*-algebras is the right counterpart of injectivity for von Neumann algebras. The classification of simple nuclear C^*-algebras is still an ongoing project, which is called the Elliott program. Being noncommutative topological spaces, C^*-algebras are expected to have classification invariants with a topological flavor. Indeed, Elliott's conjecture says that separable simple nuclear C^*-algebras in certain classes should be classified up to isomorphism by invariants coming from K-theory. Those classes for which the conjecture is verified are said to be classifiable. Thanks to the remarkable progress of the Elliott program in these two decades, there are a few known classes of classifiable

C^*-algebras with good intrinsic characterizations. Kirchberg algebras form one of such classes with the best permanence properties; for example, they are closed under taking crossed products by outer actions of countable amenable groups (see [28],[47]).

Now we ask the following question: what are plausible statements for the classification of group actions on classifiable C^*-algebras? Since amenability for groups is the same sort of property as injectivity for von Neumann algebras and nuclearity for C^*-algebras, we should keep the amenability assumption for the groups in the case of classifiable C^*-algebras too. Although finite groups clearly form the tamest subclass of amenable groups as far as analysis is concerned, it is known that K-theory for finite group actions are practically out of control. Since K-theory is the most essential element in the Elliott program, finite groups are not in a preferable situation. Indeed, recent developments of the classification of $\mathbb{Z}^2$-actions on classifiable C^*-algebras suggest that the topology of the automorphism groups of the C^*-algebras and the classifying spaces of the groups should be involved in the classification invariants of the actions, which is a completely new aspect from the case of injective factors. This would be rather natural for "noncommutative topological spaces", and would be consistent with the difficulties in finite group actions because their classifying spaces are always infinite dimensional. Trying to answer the above question, we formulate a few conjectures at the end of this note.

Throughout the note, separability or second countability for topological spaces is often assumed without mentioning it. Our standard references are [50], [51], [52] for von Neumann algebras, [2] for K-theory, and [28], [47] for the classification of nuclear C^*-algebras. The reader is referred to them for the definitions of undefined terms, and the proofs of results stated without citing references.

We end this section with recalling the basic definitions for group actions on operator algebras. Let A be a C^*-algebra or a von Neumann algebra, and let G be a locally compact group. An action α of G on A is a continuous homomorphism from G to the automorphism group $\mathrm{Aut}(A)$. We denote by $A \rtimes_\alpha G$ the crossed product of A by α, which is an analogue of a semidirect product in group theory (since we deal with only amenable groups, we do not need to distinguish the reduced crossed products from the full crossed products). We denote by A^α the fixed point subalgebra of A for α. Two actions α and β of G on A are conjugate if there exists $\gamma \in \mathrm{Aut}(A)$ satisfying $\gamma \circ \alpha_g \circ \gamma^{-1} = \beta_g$ for all $g \in G$. We denote by $U(A)$ the group of all unitaries in A. An α-cocycle u is a continuous map from G to $U(A)$ satisfying the 1-cocycle relation $u_{gh} = u_g \alpha_g(u_h)$ for all $g, h \in G$. When u is an α-cocycle, one can perturb α by u, and $\alpha^u_g = \mathrm{Ad}\, u_g \circ \alpha_g$ is again an action, where $\mathrm{Ad}\, v$ denotes the inner automorphism of A induced by $v \in U(A)$. This perturbation is often allowed for the purpose of applications because there exists an isomorphism between the crossed products by α and α^u extending the identify of A. Two actions α and β are cocycle conjugate if there exists an α-cocycle u such that β is conjugate

to α^u. A G-action α is said to be outer if α_g is not inner for any $g \in G \setminus \{e\}$. In the case of C^*-algebras, we denote by α^s the stabilization of α, which is the G-action on $A \otimes \mathbb{K}$ defined by $\alpha^s_g = \alpha_g \otimes \mathrm{Ad}\,\pi(g)$, where $\mathbb{K}$ is the set of compact operators on ℓ^2 and π is a direct sum of infinitely many copies of the regular representation of G.

2. Group Actions on Factors

2.1. Injective factors. We first recall the basics of injective factors. Factors are divided into type I, type II_1, type II_∞, and type III. A type I factor is isomorphic to $B(H)$ for a Hilbert space H, and so it is completely characterized by the dimension of H. A type II_1 factor M is characterized as an infinite dimensional factor with a finite trace τ, a linear functional satisfying $\tau(ab) = \tau(ba)$ for all $a, b \in M$ and $\tau(1) = 1$. A type II_∞ factor is isomorphic to the tensor product of a type II_1 factor and $B(\ell^2)$, and it has a unique (unbounded) semifinite trace up to a scalar multiple. Remaining factors are of type III.

We give fundamental examples of nuclear C^*-algebras and injective factors here. Let $\{n_j\}_{j=1}^\infty$ be a sequence of natural numbers greater than 1. For $k \in \mathbb{N}$, we set

$$A_k = \bigotimes_{j=1}^{k} M_{n_j}(\mathbb{C}),$$

where $M_n(\mathbb{C})$ denotes the matrix algebra over $\mathbb{C}$. Since $M_n(\mathbb{C})$ is identified with $B(\mathbb{C}^n)$, it is a C^*-algebra. We embed A_k into A_{k+1} by $\iota_k : A_k \ni x \mapsto x \otimes 1 \in A_{k+1}$. Since this embedding is isometric, the inductive limit of the system $\{A_k\}_{k\in\mathbb{N}}$ has a norm extending that of A_k. The completion A of the inductive limit with respect to this norm is called the UHF-algebra, a typical example of a simple nuclear C^*-algebra. Let τ_k be the normalized trace of A_k, which is the usual trace divided by $n_1 n_2 \cdots n_k$. Since the restriction of τ_{k+1} to A_k coincide with τ_k, there exists a trace τ of A extending τ_k. We can introduce an inner product into A by $\langle x, y\rangle = \tau(y^* x)$, and we denote by H_τ the completion of A with respect to this inner product. The UHF-algebra A acts on H_τ by left multiplication (called the GNS representation for τ), and the weak closure R of A in this representation is an injective type II_1 factor. While Murray and von Neumann showed that the isomorphism class of R does not really depend on the sequence $\{n_j\}_{j=1}^\infty$, Glimm [10] completely classified the UHF-algebras up to isomorphism whose classification invariant is the formal product of $\{n_j\}_{j=1}^\infty$, called the supernatural number. This means that there are infinitely many isomorphism classes of UHF-algebras, and one can see the sharp contrast between the two theories here. If we replace τ with other product states, we can obtain factors of other types, called Araki-Woods factors.

Type III factors are further divided into type III_λ, $0 \le \lambda \le 1$, thanks to Tomita-Takesaki theory. Every type III_λ factor with $0 < \lambda < 1$ is expressed

as the crossed product $N \rtimes_\alpha \mathbb{Z}$ with a type II_∞ factor N and a trace-scaling $\mathbb{Z}$-action α. Likewise, every type III_1 factor is expressed as $N \rtimes_\alpha \mathbb{R}$.

The classification of injective factors says that there exists a unique isomorphism class of injective factors for each type except for the type III_0 case. The injective type III_0 factors are in one-to-one correspondence with the nontransitive ergodic flows.

2.2. The classification of group actions. Although type I factors are considered as rather trivial objects in operator algebras, we nevertheless start with group actions on them in order to illustrate the difference between the two equivalence relations, conjugacy and cocycle conjugacy. It is known that every automorphism of a type I factor is inner, which shows that an action of a countable group Γ on a type I factor $B(H)$ is a synonym of a projective unitary representation of Γ in H. Therefore the classification of Γ-actions on $B(H)$ up to conjugacy is equivalent to that of projective unitary representations of Γ in H up to unitary equivalence, which cannot be accomplished even for $\Gamma = \mathbb{Z}^2$ and $H = \ell^2$. On the other hand, the set of cocycle conjugacy classes of Γ-actions on $B(\ell^2)$ is in one-to-one correspondence with the second cohomology group $H^2(\Gamma, \mathbb{T})$, which is a computable object.

To a Γ-action α on a type II_∞ factor, one can associate a homomorphism $m_\alpha : \Gamma \to \mathbb{R}_+^\times$ by the relation $\tau \circ \alpha_g = m_\alpha(g)\tau$, where τ is the unique (up to a scalar multiple) semifinite trace τ on the II_∞ factor.

Connes [5], Jones [21], and Ocneanu [42] obtained a complete classification result of countable amenable group actions on the injective type II factors. For simplicity, we state the following particular case here.

Theorem 2.1. *Let Γ be a countable amenable group.*

(1) *There exists a unique cocycle conjugacy class of outer Γ-actions on the injective type II_1 factor.*

(2) *For a given homomorphism $m : \Gamma \to \mathbb{R}_+^\times$, there exists a unique cocycle conjugacy class of outer Γ-actions α on the injective type II_∞ factor satisfying $m_\alpha = m$.*

A countable group Γ is amenable if there exists a left-invariant linear functional φ on $\ell^\infty(\Gamma)$ with $\varphi(1) = ||\varphi|| = 1$. For example, every solvable group is amenable. The amenability assumption in the above theorem is known to be necessary.

General (not necessarily outer) Γ-actions α on the injective II_1 factor are classified up to cocycle conjugacy by a relative cohomology type invariant with respect to the normal subgroup $N(\alpha) = \{g \in \Gamma |\ \alpha_g \text{ is inner}\}$. Countable amenable group actions on injective type III factors are also completely classified up to cocycle conjugacy due to Katayama, Kawahigashi, Sutherland, and Takesaki (see [23] for the final form of the statement).

While the original proofs of these results are a massive collection of case-by-case analysis depending on types, Masuda [33] recently obtained a short proof, which is independent of types. Interestingly enough, his proof is based on Evans and Kishimoto's intertwining argument ([9]) developed in C^*-algebras.

An action α of a compact group G on a factor M is said to be minimal if α is faithful as a homomorphism from G to $\mathrm{Aut}(M)$, and the relative commutant of the fixed point subalgebra $M \cap M^{\alpha\prime} = \{x \in M |\ \forall y \in M^\alpha,\ xy = yx\}$ is trivial. The following theorem is obtained by Popa and Wassermann [46] for compact Lie groups, and by Masuda and Tomatsu [34] for general compact groups.

Theorem 2.2. *For every separable compact group G, there exists a unique conjugacy class of minimal G-actions on the injective type II_1 factor and type II_∞ factor.*

As a discrete Kac algebra, the dual of a compact group is amenable. Masuda and Tomatsu actually generalized Theorem 2.1 to actions of Kac algebras, and they obtained Theorem 2.2 by a duality argument.

The full classification of minimal actions of compact groups on the injective type III factors is still in progress. See [14], [35] for related results.

2.3. Galois correspondence. The first Galois correspondence type result for group actions on operator algebras was obtained by Nakamura and Takeda [41] in 1960 for finite group actions on II_1-factors. The following generalization was obtained by Longo, Popa and the author [17].

Theorem 2.3. *Let α be a minimal action of a compact group G on a factor M. Then there exists a one-to-one correspondence between the closed subgroups $H \subset G$ and the intermediate subfactors $M^\alpha \subset N \subset M$ given by $H \mapsto M^{\alpha|_H} = N$.*

To obtain this result, more general inclusions of factors are studied in [17], which has applications to algebraic quantum field theory (see [22] for example). The crossed product inclusion of an outer action of a countable group on a factor is also a particular case.

Theorem 2.4. *Let α be an outer action of a countable group Γ on a factor M. Then there exists a one-to-one correspondence between the subgroups $\Lambda \subset \Gamma$ and the intermediate subfactors $M \subset N \subset M \rtimes_\alpha \Gamma$ given by $\Lambda \mapsto M \rtimes_\alpha \Lambda = N$.*

Tomatsu [54] generalized Theorem 2.3 to arbitrary compact quantum groups overcoming technical difficulties due to the lack of normal conditional expectations onto intermediate subfactors. Left coideals play the role of closed subgroups in this case.

2.4. Quantum group actions and Poisson boundaries. Let G be a closed subgroup of the unitary group $U(n)$, which acts on the matrix algebra $M_n(\mathbb{C})$ by conjugation. The infinite tensor product (ITP) of this action is a typical example of a minimal action of G on an injective factor. A similar construction works for a compact quantum group too if the right order of the tensor products is chosen. However, if the antipode of the quantum group is not involutive, (for example, it is the case for the q-deformations of the classical groups), the ITP action is never minimal. What is the relative commutant of the fixed point algebra then? It turns out that the relative commutant is identified with the noncommutative Poisson boundary for a convolution operator acting on the dual quantum group.

The notion of noncommutative Poisson boundaries was introduced by the author in [13] in order to answer the above question. Let M be a von Neumann algebra, and let $P : M \to M$ be a unital normal complete positive map. Although the fixed point set $H^\infty(M,P) = \{x \in M |\ P(x) = x\}$ is not necessarily an algebra, one can introduce a new product into $H^\infty(M,P)$ so that it becomes a von Neumann algebra. We call the von Neumann algebra obtained in this way the noncommutative Poisson boundary for the pair (M,P). When M is commutative, the new product of $H^\infty(M,P)$ is commutative too, and $H^\infty(M,P)$ is identified with the L^∞-space over the Poisson boundary of the random walk given by the Markov operator P.

In the case of the q-deformations of $SU(N)$, Neshveyev, Tuset and the author [19] showed that the noncommutative Poisson boundary is identified with the quantum flag manifold $SU_q(N)/\mathbb{T}^{N-1}$, which is generalized to the q-deformations of arbitrary classical groups by Tomatsu [53]. For related results on other quantum groups, see Vaes and Vander Vennet [55], [56].

3. Group Actions on C^*-algebras

3.1. K-theory. We first recall the basics of K-theory, which gives efficient isomorphism invariants of C^*-algebras. K-theory of C^*-algebras is a functor from the category of C^*-algebras to that of abelian groups, which associates two abelian groups $K_0(A)$ and $K_1(A)$ to a C^*-algebra A.

We denote by $M_n(A)$ the C^*-algebra of the n by n matrices with entries in A, by $P_n(A)$ the set of projections in $M_n(A)$, and by $U_n(A)$ the set of unitaries in $M_n(A)$. For $x \in M_m(A)$ and $y \in M_n(A)$, we set

$$x \oplus y = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix} \in M_{m+n}(A).$$

We say that two projections p, q in a C^*-algebra A are equivalent if there exists $v \in A$ such that $v^*v = p$ and $vv^* = q$. We say that p and q in $P_m(A)$ are stably equivalent if there exists $r \in P_n(A)$ such that $p \oplus r$ and $q \oplus r$ are equivalent in $M_{m+n}(A)$. Identifying $p \in P_m(A)$ with $p \oplus 0 \in P_{m+1}(A)$, we

regard $P_m(A)$ as a subset of $P_{m+1}(A)$, and we set $P_\infty(A) = \bigcup_{m=1}^\infty P_m(A)$. We denote by $K_0(A)_+$ the set of stable equivalence classes of the projections in $P_\infty(A)$, which is a semigroup with addition given by the direct sum. Its Grothendieck group is $K_0(A)$. The element in $K_0(A)$ given by the equivalence class of $p \in P_m(A)$ is denoted by $[p]_0$.

Identifying a unitary $u \in U_m(A)$ with $u \oplus 1 \in U_{m+1}(A)$, we regard $U_m(A)$ as a closed subgroup of $U_{m+1}(A)$, which induces a homomorphism from $U_m(A)/U_m(A)_0$ to $U_{m+1}(A)/U_{m+1}(A)_0$, where $U_m(A)_0$ is the connected component of the identity. The K_1-group $K_1(A)$ is defined to be the inductive limit of the system $\{U_m(A)/U_m(A)_0\}_{m=1}^\infty$. We denote by $[u]_1$ the element in $K_1(A)$ given by $u \in U_m(A)$. The quadruple $(K_0(A), K_0(A)_+, [1]_0, K_1(A))$ is an isomorphism invariant of A.

If a projection p in a C^*-algebra is equivalent to its proper subprojection, it is said to be infinite. If p is not infinite, it is said to be finite. If every projection of A is finite, the C^*-algebra A is said to be finite. If $M_n(A)$ is finite for every natural number n, the C^*-algebra A is said to be stably finite. For a stably finite A, the pair $(K_0(A), K_0(A)_+)$ is an ordered group, that is, we have $K_0(A)_+ \cap (-K_0(A)_+) = \{0\}$. II_1 factors and the UHF-algebras are typical examples of stably finite C^*-algebras.

Let A be a C^* algebra not isomorphic to the complex numbers $\mathbb{C}$. If for every $a \in A \setminus \{0\}$, there exist $x, y \in A$ satisfying $xay = 1$, the C^*-algebra A is said to be purely infinite. For a purely infinite C^*-algebra A, we have $K_0(A) = K_0(A)_+$, and one can drop $K_0(A)_+$ from the above isomorphism invariant. Type III factors are examples of purely infinite C^*-algebras.

The Cuntz algebra $\mathcal{O}_n$, for $n \geq 2$, is the universal C^*-algebra generated by isometries $S_1, S_2, \cdots, S_n$ with the relations $S_i^* S_j = \delta_{i,j}$ and $\sum_{i=1}^n S_i S_i^* = 1$. When $n = \infty$, we define $\mathcal{O}_\infty$ in a similar way imposing only the first relation. The Cuntz algebras are separable, nuclear, and purely infinite. For the K-theory of the Cuntz algebras, we have $(K_0(\mathcal{O}_n), [1]_0, K_1(\mathcal{O}_n)) \cong (\mathbb{Z}/(n-1)\mathbb{Z}, 1, \{0\})$ for finite n, and $(K_0(\mathcal{O}_\infty), [1]_0, K_1(\mathcal{O}_\infty)) \cong (\mathbb{Z}, 1, \{0\})$.

Kasparov's KK-theory is a functor associating two abelian groups $KK^0(A, B)$ and $KK^1(A, B)$ to C^*-algebras A and B, which are contravariant for A and covariant for B. The group $KK^0(A, B)$ is often simply denoted by $KK(A, B)$. We have $KK^*(\mathbb{C}, B) = K_*(B)$, and $KK^*(A, \mathbb{C}) = K^*(A)$, the K-homology group of A. Every homomorphism ρ from A to B gives a KK-class $KK(\rho) \in KK(A, B)$. The most remarkable feature of KK-theory is the existence of an associated product, called the Kasparov product, which is a generalization of the composition of homomorphisms:

$$KK^i(A, B) \times KK^j(B, C) \ni (x, y) \mapsto x\#y \in KK^{i+j}(A, C),$$

where $KK^{*+2}(A, B) = KK^*(A, B)$. When there exist $x \in KK(A, B)$ and $y \in KK(B, A)$ satisfying $x\#y = KK(\mathrm{id}_A)$ and $y\#x = KK(\mathrm{id}_B)$, the two C^*-algebras A and B are said to be KK-equivalent.

A C^*-algebra that is KK-equivalent to an abelian C^*-algebra is said to be K-abelian. We denote by $\mathcal{N}$ the category of separable K-abelian C^*-algebras, and call $\mathcal{N}$ the UCT class. It is an important open problem whether every nuclear C^*-algebra belongs to $\mathcal{N}$. We denote by $\mathcal{N}_{\rm nuc}$ the set of nuclear C^*-algebras in $\mathcal{N}$. The UHF-algebras and the Cuntz algebras are in $\mathcal{N}_{\rm nuc}$.

3.2. Classifiable C^*-algebras. In the early stage, the classification of nuclear C^*-algebras developed by extending the classes of building blocks of inductive limit C^*-algebras. When a C^*-algebra A has an increasing sequence of C^*-subalgebras $\{A_k\}_{k=1}^{\infty}$ whose union is dense in A, we say that A is the inductive limit of $\{A_k\}_{k=1}^{\infty}$. If each A_k is a direct sum of $A_{k,j}$, $1 \le j \le m_k$, we say that $A_{k,j}$ is a building block of A. There are several important classes of inductive limit C^*-algebras specified by the classes of building blocks. For example, if every building block is a matrix algebra, we say that A is an AF-algebra. The UHF-algebras are inductive limits of full matrix algebras, and they form a subclass of AF-algebras. More generally, if every building block is of the form $pM_n(C(\Omega))p$, where Ω is a compact Hausdorff space and p is a projection in $M_n(C(\Omega))$, the C^*-algebra A is said to be an AH-algebra. AT-algebras, for which Ω is the circle $\mathbb{T} = \{z \in \mathbb{C} \ \ |z| = 1\}$, form a subclass of AH-algebras. If every building block is a subalgebra of $M_n(C(\Omega))$, the C^*-algebra A is said to be an ASH-algebra.

After Glimm's classification of the UHF-algebras, Bratteli and Elliott classified AF-algebras in the 70s. In modern terms, Elliott's classification invariant is the triple $(K_0(A), K_0(A)_+, [1]_0)$ (the K_1-group is trivial for an AF-algebra). In the early 90s, Elliott generalized this result to AT-algebras of real rank 0, where the real rank of a C^*-algebra is a generalization of the covering dimension of a topological space. This is the breakthrough of the remarkable developments of the classification of nuclear C^*-algebras in these two decades.

The first classifiable class of nuclear C^*-algebras without referring to inductive limits was discovered in the purely infinite case. A separable simple nuclear purely infinite C^*-algebra is said to be a Kirchberg algebra. The Cuntz algebras are fundamental examples of Kirchberg algebras. Kirchberg obtained the following result in the mid 90s.

Theorem 3.1. *Let A be a unital nuclear separable simple C^*-algebra. Then the following hold:*

(1) *The tensor product $A \otimes \mathcal{O}_2$ is isomorphic to $\mathcal{O}_2$.*

(2) *If A is purely infinite, the tensor product $A \otimes \mathcal{O}_\infty$ is isomorphic to A.*

The above theorem shows that $\mathcal{O}_2$ plays the role of a zero element, and $\mathcal{O}_\infty$ plays the role of a unit element for tensor product. This fits well with the facts that $\mathcal{O}_2$ is KK-equivalent to $\{0\}$, and that $\mathcal{O}_\infty$ is KK-equivalent to $\mathbb{C}$.

Based on Theorem 3.1, Kirchberg and Phillips showed that the KK-theory of Kirchberg algebras is given by the asymptotically unitary equivalence classes of homomorphisms, and they obtained the following classification theorem.

Theorem 3.2. *Let A and B be unital Kirchberg algebras.*

(1) *The two C^*-algebras A and B are KK-equivalent if and only if they are stably isomorphic, that is, $A \otimes \mathbb{K} \cong B \otimes \mathbb{K}$.*

(2) *Assume that $A, B \in \mathcal{N}_{\mathrm{nuc}}$. Then A and B are isomorphic if and only if*

$$(K_0(A), [1_A]_0, K_1(A)) \cong (K_0(B), [1_B]_0, K_1(B)).$$

For any countable abelian groups M_0 and M_1, and any $m \in M_0$, there exists a model of A satisfying the assumption of (2) such that

$$(M_0, m, M_1) \cong (K_0(A), [1_A]_0, K_1(A)).$$

The Elliott program in the stably finite case is still in progress. The first classifiable class of stably finite C^*-algebras without referring to inductive limits was provided by Huaxin Lin. A unital simple C^*-algebra A is said to have tracial topological rank 0 if it satisfies the following condition: for every $\varepsilon > 0$, for every $a \in A_+ \setminus \{0\}$, and for every finite set $F \subset A$, there exists a non-zero projection $p \in A$ satisfying the following:

(1) For every $x \in F$, we have $\|px - xp\| < \varepsilon$.

(2) The projection $1 - p$ is equivalent to a projection in the closure of aAa.

(3) There exists a finite dimensional C^*-subalgebra B of pAp and its finite subset G such that the distance between G and $\{pxp \in A;\ x \in F\}$ is less than ε.

When one expresses $x \in F$ as a matrix

$$x = \begin{pmatrix} (1-p)x(1-p) & (1-p)xp \\ px(1-p) & pxp \end{pmatrix},$$

the condition (1) means that the off diagonal entries are small in norm, and the condition (2) means that the left up corner is small in the order of projections. If p can be taken to be 1, the condition (3) is nothing but the local characterization of AF-algebras (see [8]). Simple AF-algebras and simple AT-algebras of real rank 0 have tracial topological rank 0.

Lin [29] proved the following classification theorem.

Theorem 3.3. *Let A and B be unital simple C^*-algebras in $\mathcal{N}_{\mathrm{nuc}}$ with tracial topological rank 0. Then A and B are isomorphic if and only if*

$$(K_0(A), K_0(A)_+, [1_A]_0, K_1(A)) \cong (K_0(B), K_0(B)_+, [1_B]_0, K_1(B)).$$

An abstract characterization of the quadruple $(K_0(A), K_0(A)_+, [1_A]_0, K_1(A))$ in Theorem 3.3 is also known. A similar result holds in the case of tracial topological rank 1, where the set of traces are included in the classification invariant (see [30]).

Jiang and Su [20] constructed a unital simple nuclear ASH-algebra $\mathcal{Z}$ without nontrivial projections that is KK-equivalent to the complex numbers $\mathbb{C}$. One can regard $\mathcal{Z}$ as a stably finite version of $\mathcal{O}_\infty$. Every C^*-algebra in known classifiable classes absorbs $\mathcal{Z}$ by tensor product, which is believed to be the key property for the classification of nuclear C^*-algebras. For the latest classification results along this line, the reader is referred to Lin [31], Lin and Niu [32], and Winter [57].

3.3. The Rohlin property. The Rohlin property, originally coming from ergodic theory, was first used in operator algebras in Connes's classification of cyclic group actions of injective type II factors. In the 90s, its importance drew attention of specialists in C^*-algebras, and systematic analysis of it was launched. The reader is referred to [12] for the developments of the subject up to 2000, mainly due to Kishimoto.

We say that an automorphism α of a unital C^*-algebra A has the Rohlin property if for any natural number n, any $\varepsilon > 0$, and any finite set $F \subset A$, there exists a partition of unity consisting of projections $\{e_i\}_{i=0}^{n-1} \cup \{f_i\}_{i=0}^{n} \subset A$ satisfying

$$\|\alpha(e_i) - e_{i+1}\| < \varepsilon, \quad 0 \le \forall i \le n-2,$$

$$\|\alpha(f_i) - f_{i+1}\| < \varepsilon, \quad 0 \le \forall i \le n-1,$$

$$\|xe_i - e_i x\| < \varepsilon, \quad 0 \le \forall i \le n-1,\ \forall x \in F,$$

$$\|xf_i - f_i x\| < \varepsilon, \quad 0 \le \forall i \le n,\ \forall x \in F.$$

In what follows, we often identify a single automorphism with the $\mathbb{Z}$-action generated by it. An automorphism is said to be aperiodic if the corresponding $\mathbb{Z}$-action is outer. For the classification of automorphisms α of a C^*-algebra A up to cocycle conjugacy, the most important property is the stability of α, which is a statement of the following type: every unitary $u \in A$ in a certain class can be approximated by $v\alpha(v^*)$ with v in the same class. For α to have this property, it needs to be outer in a very strong sense. The Rohlin property is a means to deduce the stability of α (see [12] for details).

Except for inductive limit actions, the first classification result of group actions on C^*-algebras was obtained by Kishimoto [25]. For a C^*-algebra A with a unique trace τ, and for $\alpha \in \mathrm{Aut}(A)$, we denote by $\overline{\alpha}$ the weakly continuous extension of α to the weak closure of A in the GNS representation for τ. We say that α is strongly outer if $\overline{\alpha}$ is outer.

Theorem 3.4. *Let A be a UHF-algebra, and let $\alpha \in \mathrm{Aut}(A)$. The following conditions are equivalent:*

(1) *α has the Rohlin property.*

(2) *α^n is strongly outer for any $n \in \mathbb{Z} \setminus \{0\}$.*

(3) *The crossed product $A \rtimes_\alpha \mathbb{Z}$ has a unique trace.*

(4) *The crossed product $A \rtimes_\alpha \mathbb{Z}$ is a simple AT-algebra of real rank 0.*

Moreover, there exists a unique cocycle conjugacy class of automorphisms with the Rohlin property.

The condition (2) is a useful criterion to see if a given automorphism has the Rohlin property. In fact, it is easy to construct an aperiodic automorphism without satisfying this condition. The condition (4) means that $A \rtimes_\alpha \mathbb{Z}$ is still in a classifiable class, which suggests that sufficiently large classifiable classes should have a permanence property under crossed products by $\mathbb{Z}$-actions with the Rohlin property. Several authors ([26], [27], [37], [38], [43]) have generalized various aspects of Theorem 3.4 to simple AT-algebras of real rank 0 and more generally, simple C^*-algebras of tracial rank 0. For the Jiang-Su algebra, Sato [48] showed that there exists a unique cocycle conjugacy class of strongly outer $\mathbb{Z}$-actions.

Following Kishimoto's strategy, Nakamura [40] completely classified aperiodic automorphisms of Kirchberg algebras. The Rohlin property is automatic in this case.

Theorem 3.5. *Every aperiodic automorphism of a unital Kirchberg algebra A has the Rohlin property. Moreover, the following conditions are equivalent for two aperiodic automorphisms $\alpha, \beta \in \mathrm{Aut}(A)$:*

(1) *$KK(\alpha) = KK(\beta)$,*

(2) *there exist $\gamma \in \mathrm{Aut}(A)$ and $u \in U(A)$ satisfying $KK(\gamma) = KK(\mathrm{id})$ and $\beta = \mathrm{Ad}\, u \circ \gamma \circ \alpha \circ \gamma^{-1}$.*

To generalize the above results to group actions, we need to formulate the Rohlin property for group actions first. Nakamura [39] and the author [15] discussed it for $\mathbb{Z}^N$-actions and finite group actions respectively. We present it in a unified form here.

Let Γ be a countable group. We say that a Γ-action α on a unital C^*-algebra A has the Rohlin property if for any finite set $\Gamma_0 \subset \Gamma \setminus \{e\}$, there exist finitely many subgroups $\Lambda_1, \Lambda_2, \ldots, \Lambda_r < \Gamma$ of finite index such that Γ_0 does not intersect with any conjugate of Λ_j for $1 \leq j \leq r$, and the following holds: for any $\varepsilon > 0$, any finite set $F \subset A$, and any finite set $\Gamma_1 \subset \Gamma$, there exists a

partition of unity consisting of projections $\bigcup_{j=1}^{r}\{e_k^{(j)}\}_{k\in\Gamma/\Lambda_j}\subset A$ satisfying

$$\left\|\alpha_g(e_k^{(j)})-e_{gk}^{(j)}\right\|<\varepsilon,\quad 1\le\forall j\le r,\ \forall k\in\Gamma/\Lambda_j,\ \forall g\in\Gamma_1,$$

$$\left\|xe_k^{(j)}-e_k^{(j)}x\right\|<\varepsilon,\quad 1\le\forall j\le r,\ \forall k\in\Gamma/\Lambda_j,\ \forall x\in F.$$

It is easy to see that this condition forces Γ to be residually finite.

Toward the classification of Γ-actions, we have to take the following two steps: (1) to show that every (strongly, if A is stably finite) outer Γ-action has the Rohlin property, (2) to classify Γ-actions with the Rohlin property up to cocycle conjugacy. When Γ is finite, the Rohlin property is reduced to the condition with $r=1$, $\Lambda_j=\{e\}$, and it gives a strong K-theoretical constraint. In fact, there are many K-theoretical obstructions for the step (1), and we have to give up general classification as in the two theorems above. On the other hand, the step (2) has already been done by the author. For $\Gamma=\mathbb{Z}^N$, it seems, at least to the author, that there is no obstruction to these two steps. We report on the recent progress of these cases in the next two subsections.

3.4. Finite group actions. The reader is referred to [15],[16] for the proofs of the results stated in this subsection.

Let Γ be a finite group. For a Γ-action α on a C^*-algebra A, the Rohlin property takes the following form: for every $\varepsilon>0$ and every finite set $F\subset A$, there exists a partition of unity consisting of projections $\{e_g\}_{g\in\Gamma}$ in A satisfying

$$\|\alpha_g(e_h)-e_{gh}\|<\varepsilon,\quad\forall g,h\in\Gamma,$$

$$\|xe_g-e_gx\|<\varepsilon,\quad\forall g\in\Gamma,\ \forall x\in F.$$

This condition implies that the following equation holds in $K_0(A)$ if ε is sufficiently small:

$$\sum_{g\in\Gamma}K_0(\alpha_g)([e_e]_0)=[1]_0,$$

which looks a strong constraint for $K_0(A)$ as a Γ-module. In fact, a much stronger statement holds. We say that a Γ-module M is cohomologically trivial if the Tate cohomology $\widehat{H}^*(\Lambda,M)$ vanishes for every subgroup Λ of Γ (see [4]). If nM is cohomologically trivial for all $n\in\mathbb{N}$, we say that M is completely cohomologically trivial.

Theorem 3.6. *Let α be an action of a finite group Γ on a simple unital C^*-algebra A. If α has the Rohlin property, then $K_0(A)$ and $K_1(A)$ are completely cohomologically trivial Γ-modules.*

This immediately implies that any C^*-algebra A with either $K_0(A)\cong\mathbb{Z}$ or $K_1(A)\cong\mathbb{Z}$, e.g. $A=\mathcal{O}_\infty$, has no nontrivial finite group action with the Rohlin property.

Although there is little hope to classify general outer actions of finite groups on Kirchberg algebras, we have the following theorem for those with the Rohlin property.

Theorem 3.7. *Let Γ be a finite group.*

(1) *Let A be a unital Kirchberg algebra in $\mathcal{N}_{\mathrm{nuc}}$. If α and β are Γ-actions on A with the Rohlin property such that $K_*(\alpha_g) = K_*(\beta_g)$ for all $g \in \Gamma$, then there exists $\theta \in \mathrm{Aut}(A)$ satisfying $K_*(\theta) = 1$ and $\theta \circ \alpha_g \circ \theta^{-1} = \beta_g$ for all $g \in \Gamma$.*

(2) *For countable completely cohomologically trivial Γ-modules M_0 and M_1, there exists a Γ-action α with the Rohlin property on a unital Kirchberg algebra A in $\mathcal{N}_{\mathrm{nuc}}$ such that $K_i(A)$ is isomorphic to M_i as a Γ-module for $i = 0, 1$.*

The statement (1) holds for A in the class in Theorem 3.3 too.

The Rohlin property for finite group actions is also useful to formulate the following Γ-equivariant version of Theorem 3.1,(1).

Theorem 3.8. *Let α be an outer action of a finite group Γ on a separable simple unital nuclear C^*-algebra A. Then the Γ-action $\mathrm{id} \otimes \alpha$ on $\mathcal{O}_2 \otimes A$ is conjugate to a unique (up to conjugate) Γ-action on $\mathcal{O}_2$ with the Rohlin property.*

If we restrict ourself to $\mathbb{Z}/2\mathbb{Z}$-actions on $\mathcal{O}_2$, we have a reasonable classification result. We say that a $\mathbb{Z}/2\mathbb{Z}$-action α on a C^*-algebra A is strongly approximately inner if there exists a sequence of unitaries $\{u_n\}_{n=1}^{\infty} \subset A^{\alpha}$ such that the sequence $\{u_n x u_n^*\}_{n=1}^{\infty}$ converges to $\alpha_1(x)$ for all $x \in A$. If moreover we can choose u_n to be self-adjoint, we say that α is approximately representable. It is easy to see that the dual action of an approximately representable action has the Rohlin property. Showing that strongly approximate innerness implies approximate representability in the case of $\mathcal{O}_2$, the author obtained the following theorem by classifying the dual actions.

Theorem 3.9. *Let α and β be outer strongly approximately inner $\mathbb{Z}/2\mathbb{Z}$-actions on $\mathcal{O}_2$.*

(1) *Two actions α and β are cocycle conjugate if and only if their crossed products are isomorphic.*

(2) *Two actions α and β are conjugate if and only if their fixed point algebras are isomorphic.*

The K-groups of the crossed product of $\mathcal{O}_2$ by any $\mathbb{Z}/2\mathbb{Z}$-action are always uniquely 2-divisible (i.e. multiplying by 2 is a group automorphism). On the other hand, for any countable uniquely 2-divisible abelian groups M_0, M_1, there exists an outer $\mathbb{Z}/2\mathbb{Z}$-action α on $\mathcal{O}_2$ satisfying $K_i(\mathcal{O}_2 \rtimes_{\alpha} \mathbb{Z}/2\mathbb{Z}) \cong M_i$ for $i = 0, 1$. Every known $\mathbb{Z}/2\mathbb{Z}$-action on $\mathcal{O}_2$ is strongly approximately inner (and hence approximately representable).

The reader is referred to [38], [44], [45] for the permanence property of classifiable classes under the crossed products by finite group actions with the Rohlin property (or its variant).

3.5. $\mathbb{Z}^N$-actions. The Rohlin property for $\mathbb{Z}^N$-actions on C^*-algebras was first discussed by Nakamura [39]. He showed that the Rohlin property of $\mathbb{Z}^2$-actions on the UHF-algebras is equivalent to strong outerness as in Theorem 3.4, and he classified product type $\mathbb{Z}^2$-actions with the Rohlin property. This classification result was generalized by Katsura and Matui [24] to general $\mathbb{Z}^2$-actions with the Rohlin property on the UHF-algebras (see also [37]).

Matui and the author [18] recently classified a large class of outer $\mathbb{Z}^2$-actions on a Kirchberg algebra A by $KK^1(A,A)$. We say that two actions α and β of a group Γ on A are KK-trivially cocycle conjugate if there exist $\gamma \in \mathrm{Aut}(A)$ with $KK(\gamma) = KK(\mathrm{id})$ and α-cocycle u satisfying $\gamma \circ \beta_g \circ \gamma^{-1} = \alpha^u_g$ for all $g \in \Gamma$.

Theorem 3.10. *Let A be a unital Kirchberg algebra.*

(1) *Every outer $\mathbb{Z}^N$-action on A has the Rohlin property.*

(2) *There exists a one-to-one correspondence between*

$$\{x \in KK^1(A,A) |\ [1]_0 \# x = 0 \in K_1(A)\},$$

and the set of KK-trivially cocycle conjugacy classes of outer $\mathbb{Z}^2$-actions α with $KK(\alpha_g) = KK(\mathrm{id})$ for all $g \in \mathbb{Z}^2$.

We can see from (2) that there exist exactly $n-1$ cocycle conjugacy classes of outer $\mathbb{Z}^2$-actions on the Cuntz algebra $\mathcal{O}_n$ for finite n.

The classification invariant in $KK^1(A,A)$ arises in the following way. Let G be a path connected topological group, and let $g, h \in G$ with $gh = hg$. We choose a continuous path $\{g(t)\}_{t\in[0,1]}$ in G connecting e and g. Then the path $\{g(t)hg(t)^{-1}h^{-1}\}_{t\in[0,1]}$ is a loop in G. It is easy to show that the class of the loop in the fundamental group $\pi_1(G)$ does not really depend on the choice of the path $\{g(t)\}_{t\in[0,1]}$. For a $\mathbb{Z}^2$-action α, we could apply this argument to $G = \mathrm{Aut}(A)_0$, the connected component of id, and the images of the canonical generators of $\mathbb{Z}^2$ if they were in $\mathrm{Aut}(A)_0$. Although our assumption $KK(\alpha_g) = KK(\mathrm{id})$ does not really imply $\alpha_g \in \mathrm{Aut}(A)_0$, it is known that $\theta \in \mathrm{Aut}(A)$ satisfies $KK(\theta) = KK(\mathrm{id})$ if and only if $\theta \otimes \mathrm{id} \in \mathrm{Aut}(A \otimes \mathbb{K})_0$. Thus we can apply the argument to the stabilization of α. On the other hand, Dadarlat [6] showed that $\pi_1(\mathrm{Aut}(A \otimes \mathbb{K})_0)$ is isomorphic to $KK^1(A,A)$, and we get an invariant of α in $KK^1(A,A)$.

For $\mathbb{Z}^N$-actions, Matui [36], and Matui and the author [18] obtained the following uniqueness result.

Theorem 3.11. *Let A be either $\mathcal{O}_2$, $\mathcal{O}_\infty$, or $\mathcal{O}_\infty \otimes B$ with B being a UHF-algebra of infinite type (i.e. $B \cong B \otimes B$). Then there exists a unique cocycle conjugacy class of outer $\mathbb{Z}^N$-actions on A for any natural number N.*

A unital C^*-algebra $A \neq \mathbb{C}$ is said to be strongly self-absorbing if there exists an isomorphism ρ from A onto $A \otimes A$ that is approximately unitarily equivalent to the inclusion map $A \ni x \mapsto x \otimes 1 \in A \otimes A$. A C^*-algebra A is said to be K_1-injective if the canonical map from $U(A)/U(A)_0$ to $K_1(A)$ is injective. The C^*-algebras in the statement of Theorem 3.11 are examples of strongly self-absorbing K_1-injective C^*-algebras. The UHF-algebras of infinite type and the Jiang-Su algebra $\mathcal{Z}$ are other examples. Indeed, Katsura and Matui [24], and Matui and Sato [38] showed the following.

Theorem 3.12. *Let A be either a UHF-algebra of infinite type or the Jiang-Su algebra $\mathcal{Z}$. Then there exists a unique cocycle conjugacy class of strongly outer $\mathbb{Z}^2$-actions on A.*

3.6. Conjectures. Before ending this note, we clarify what our classification invariant in Theorem 3.10 means in obstruction theory (see [49] for example), and present two conjectures, which would generalize Theorem 3.10, Theorem 3.11, and Theorem 3.12.

Let Γ be a discrete group, and let G be a topological group. We denote by $B\Gamma$ the classifying space of Γ, and denote by $E\Gamma$ the universal covering space of $B\Gamma$. To a homomorphism $\rho : \Gamma \to G$, we can associate a principal G-bundle $\mathcal{P}_\rho$ over $B\Gamma$, which is the quotient space of $E\Gamma \times G$ by the equivalence relation $(x \cdot \gamma, g) \sim (x, \rho(\gamma)g)$ for $x \in E\Gamma$, $g \in G$ and $\gamma \in \Gamma$. For two homomorphisms ρ and σ, whether $\mathcal{P}_\rho$ and $\mathcal{P}_\sigma$ are isomorphic or not can be determined as follows. Let $\mathcal{I}_{\rho,\sigma}$ be the quotient space of $E\Gamma \times G$ by the equivalence relation $(x \cdot \gamma, g) \sim (x, \rho(\gamma)g\sigma(\gamma)^{-1})$ for $x \in E\Gamma$, $g \in G$ and $\gamma \in \Gamma$, which is a fiber bundle over $B\Gamma$. Then the two principal G-bundles $\mathcal{P}_\rho$ and $\mathcal{P}_\sigma$ over $B\Gamma$ are isomorphic if and only if $\mathcal{I}_{\rho,\sigma}$ has a continuous section.

Assume now that $\Gamma = \mathbb{Z}^2$ and G is path connected. For $g = \rho((1,0))$ and $h = \rho((0,1))$, the π_1-class of the loop $\{g(t)hg(t)^{-1}h^{-1}\}_{t\in[0,1]}$ discussed in the previous subsection can be identified with the the primary obstruction class in

$$H^2(B\mathbb{Z}^2, \pi_1(G)) = H^2(\mathbb{T}^2, \pi_1(G)) \cong \pi_1(G),$$

for the existence of a continuous section of the principal G-bundle $\mathcal{P}_\rho$ over $\mathbb{T}^2$. The author would like to thank Sergey Neshveyev for this observation.

Whenever an action α of a group Γ on a C^*-algebra A is given, we can associate to α a principal $\mathrm{Aut}(A)$-bundle $\mathcal{P}_\alpha$ over the classifying space $B\Gamma$, where $\mathrm{Aut}(A)$ is not necessarily connected. We denote by $\mathrm{Aut}(A)_s$ the subgroup of $\mathrm{Aut}(A \otimes \mathbb{K})$ generated by $\mathrm{Aut}(A)$ and the inner automorphism group of $A \otimes \mathbb{K}$, where we identifying $\theta \in \mathrm{Aut}(A)$ with $\theta \otimes \mathrm{id} \in \mathrm{Aut}(A \otimes \mathbb{K})$. We regard the stabilization α^s of α as a homomorphism from Γ to $\mathrm{Aut}(A)_s$, and denote by $\mathcal{P}^s_\alpha$ the corresponding principal $\mathrm{Aut}(A)_s$-bundle over $B\Gamma$. If two Γ-actions α and β are cocycle conjugate, their stabilizations are conjugate in $\mathrm{Aut}(A)_s$, and so the two principal $\mathrm{Aut}(A)_s$-bundles $\mathcal{P}^s_\alpha$ and $\mathcal{P}^s_\beta$ are isomorphic. When A is a unital Kirchberg algebra, and $KK(\alpha_g) = KK(\mathrm{id})$ for all $g \in \Gamma$, we can regard

α^s as a homomorphism from Γ to $\mathrm{Aut}(A\otimes\mathbb{K})_0$ too, and we denote by $\mathcal{P}_\alpha^{s,0}$ the corresponding principal $\mathrm{Aut}(A\otimes\mathbb{K})_0$-bundle over $B\Gamma$.

Conjecture 1. *Let A be a unital Kirchberg algebra, and let Γ be a countable amenable group whose classifying space $B\Gamma$ has the homotopy type of a finite CW complex.*

(1) *Two outer actions α and β of Γ on A are cocycle conjugate if and only if the principal $\mathrm{Aut}(A)_s$-bundles $\mathcal{P}_\alpha^s$ and $\mathcal{P}_\beta^s$ over $B\Gamma$ are isomorphic.*

(2) *Let α and β be outer actions of Γ on A satisfying $KK(\alpha_g)=KK(\beta_g)=KK(\mathrm{id})$ for all $g\in\Gamma$. The two actions α and β are KK-trivially cocycle conjugate if and only if the principal $\mathrm{Aut}(A\otimes\mathbb{K})_0$-bundles $\mathcal{P}_\alpha^{s,0}$ and $\mathcal{P}_\beta^{s,0}$ over $B\Gamma$ are isomorphic.*

Conjecture 1 is true for $\Gamma=\mathbb{Z}$ thanks to Theorem 3.5, and (2) is true for $\Gamma=\mathbb{Z}^2$ thanks to Theorem 3.10. Finite groups are excluded from the conjecture because the classifying space $B\Gamma$ does not have the homotopy type of a finite CW complex for any nontrivial finite group Γ.

Classical obstruction theory says that whether $\mathcal{P}_\alpha^{s,0}$ and $\mathcal{P}_\beta^{s,0}$ are isomorphic or not can be determined by computing relevant cohomology classes in

$$H^n(B\Gamma,\pi_{n-1}(\mathrm{Aut}(A\otimes\mathbb{K})_0)),\quad 2\le n\le\dim B\Gamma.$$

This can be done, at least in principle, because Dadarlat [8] computed the homotopy groups $\pi_n(\mathrm{Aut}(A\otimes\mathbb{K}))$ for the Kirchberg algebras.

Dadarlat and Winter [7] showed that if A is a strongly self-absorbing K_1-injective C^*-algebra, the homotopy groups $\pi_n(\mathrm{Aut}(A))$ are trivial for $n\ge 0$. This implies that if α is a Γ-action on such a C^*-algebra A, the principal $\mathrm{Aut}(A)$-bundle $\mathcal{P}_\alpha$ over $B\Gamma$ is trivial.

Conjecture 2. *Let Γ be a countable amenable group whose classifying space $B\Gamma$ has the homotopy type of a finite CW complex.*

(1) *If A is either $\mathcal{O}_2$, $\mathcal{O}_\infty$ or $\mathcal{O}_\infty\otimes B$ with B being a UHF-algebra of infinite type, there exists a unique cocycle conjugacy class of outer Γ-actions on A.*

(2) *If A is either a UHF-algebra of infinite type or the Jiang-Su algebra $\mathcal{Z}$, there exists a unique cocycle conjugacy class of strongly outer Γ-actions on A.*

If Γ is a cocompact lattice of a simply connected solvable Lie group S, we may choose S for $E\Gamma$ because S is homeomorphic to $\mathbb{R}^n$. Thus the assumption on Γ in the two conjectures above is satisfied. For a Γ-action α on a C^*-algebra A, we let

$$M_\alpha=\{f\in C^b(S,A)|\ f(x\gamma)={\alpha_\gamma}^{-1}(f(x)),\ \forall x\in S,\ \gamma\in\Gamma\},$$

where $C^b(S, A)$ is the set of bounded continuous maps from S to A. The C^*-algebra M_α is identified with the set of continuous sections of the fiber bundle $\mathcal{P}_\alpha \times_{\mathrm{Aut}(A)} A$ over $B\Gamma$ associated with $\mathcal{P}_\alpha$. Therefore the isomorphism class of $\mathcal{P}_\alpha$ determines M_α, and hence the K-theory of M_α. On the other hand, since the crossed product $M_\alpha \rtimes_\lambda S$ by the left translation action λ is stably isomorphic to the crossed product $A \rtimes_\alpha \Gamma$, the K-theory of M_α is the same as that of $A \rtimes_\alpha \Gamma$, up to degree change, thanks to repeated use of Connes's Thom isomorphism. This means that the isomorphism class of $\mathcal{P}_\alpha$ determines the K-theory of $A \rtimes_\alpha \Gamma$, which is consistent with the two conjectures above.

Acknowledgements

The author would like to thank Hiroki Matui for his critical reading of the first draft, and Mariko Izumi for everything.

References

[1] H. Araki, *Mathematical theory of quantum fields*, Translated from the 1993 Japanese original by Ursula Carow-Watamura. International Series of Monographs on Physics, 101. Oxford University Press, New York, 1999.

[2] B. Blackadar, *K-theory for operator algebras*, Mathematical Sciences Research Institute Publications, 5. Cambridge University Press, Cambridge, 1998.

[3] O. Bratteli and D. W. Robinson, *Operator algebras and quantum statistical mechanics. 1. C^*- and W^*-algebras, symmetry groups, decomposition of states,* Second edition. Texts and Monographs in Physics. Springer-Verlag, New York, 1987.

[4] K. S. Brown, *Cohomology of groups,* Graduate Texts in Mathematics, 87. Springer-Verlag, New York-Berlin, 1982.

[5] A. Connes, *Outer conjugacy classes of automorphisms of factors.* Ann. Sci. École Norm. Sup. (4) **8** (1975), 383–419.

[6] M. Dadarlat, *The homotopy groups of the automorphism group of Kirchberg algebras*, J. Noncommut. Geom. **1** (2007), 113–139.

[7] M. Dadarlat and W. Winter, *On the KK-theory of strongly self-absorbing C^*-algebras*, Math. Scand. **104** (2009), 95–107.

[8] K. R. Davidson, *C^*-algebras by Example*, Fields Institute Monographs, Amer. Math. Soc., Providence, R.I., 1996.

[9] D. Evans and A. Kishimoto, *Trace scaling automorphisms of certain stable AF algebras*, Hokkaido Math. J. **26** (1997), 211–224.

[10] J. G. Glimm, *On a certain class of operator algebras*, Trans. Amer. Math. Soc. **95** (1960), 318–340.

[11] R. Haag, *Local quantum physics. Fields, particles, algebras,* Second edition. Texts and Monographs in Physics. Springer-Verlag, Berlin, 1996.

[12] M. Izumi, *The Rohlin property for automorphisms of C^*-algebras*, Mathematical Physics in Mathematics and Physics (Siena, 2000), 191–206, Fields Inst. Commun., **30**, Amer. Math. Soc., Providence, RI, 2001.

[13] M. Izumi, *Non-commutative Poisson boundaries and compact quantum group actions*, Adv. Math. **169** (2002), 1–57.

[14] M. Izumi, *Canonical extension of endomorphisms of type III factors*, Amer. J. Math. **125** (2003), 1–56.

[15] M. Izumi, *Finite group actions on C^*-algebras with the Rohlin property. I*, Duke Math. J. **122** (2004), 233–280.

[16] M. Izumi, *Finite group actions on C^*-algebras with the Rohlin property. II*, Adv. Math. **184** (2004), 119–160.

[17] M. Izumi, R. Longo and S. Popa, *A Galois correspondence for compact groups of automorphisms of von Neumann algebras with a generalization to Kac algebras*, J. Funct. Anal. **155** (1998), 25–63.

[18] M. Izumi and H. Matui, *$\mathbb{Z}^2$-actions on Kirchberg algebras*, to appear in Adv. Math. arXiv:0902.0194.

[19] M. Izumi, S. Neshveyev and L. Tuset, *Poisson boundary of the dual of* $\mathrm{SU}_q(n)$, Comm. Math. Phys. **262** (2006), 505–531.

[20] X. Jiang and H. Su, *On a simple unital projectionless C^*-algebra*, Amer. J. Math. **121** (1999), 359–413.

[21] V. F. R. Jones, *Actions of finite groups on the hyperfinite type* II_1 *factor*. Mem. Amer. Math. Soc. **28** (1980), no. 237.

[22] Y. Kawahigashi and R. Longo, *Classification of local conformal nets. Case $c < 1$*, Ann. of Math. (2) **160** (2004), 493–522.

[23] Y. Katayama, C. E. Sutherland and M. Takesaki, *The characteristic square of a factor and the cocycle conjugacy of discrete group actions on factors,* Invent. Math. **132** (1998), 331–380.

[24] T. Katsura and H. Matui, *Classification of uniformly outer actions of $\mathbb{Z}^2$ on UHF algebras*, Adv. Math. **218** (2008), 940–968.

[25] A. Kishimoto, *The Rohlin property for automorphisms of UHF algebras*. J. Reine Angew. Math. **465** (1995), 183–196.

[26] A. Kishimoto, *Automorphisms of AT algebras with the Rohlin property*, J. Operator Theory **40** (1998), 277–294.

[27] A. Kishimoto, *Unbounded derivations in AT algebras*, J. Funct. Anal. **160** (1998), 270–311.

[28] H. Lin, *An introduction to the classification of amenable C^*-algebras*, World Scientific Publishing Co., Inc., River Edge, NJ, 2001.

[29] H. Lin, *Classification of simple C^*-algebras of tracial topological rank zero*, Duke Math. J. **125** (2004), 91–119.

[30] H. Lin, *Simple nuclear C^*-algebras of tracial topological rank one*, J. Funct. Anal. **251** (2007), 601–679.

[31] H. Lin, *Asymptotically unitary equivalence and classification of simple amenable C^*-algebras*, preprint, arXiv:0806.0636.

[32] H. Lin, and Z. Niu, *Lifting KK-elements, asymptotic unitary equivalence and classification of simple C^*-algebras*, Adv. Math. **219** (2008), 1729–1769.

[33] T. Masuda, *Unified approach to classification of actions of discrete amenable groups on injective factors*, preprint.

[34] T. Masuda and R. Tomatsu, *Classification of minimal actions of a compact Kac algebra with amenable dual*, Comm. Math. Phys. **274** (2007), 487–551.

[35] T. Masuda and R. Tomatsu, *Classification of minimal actions of a compact Kac algebra with amenable dual on injective factors of type III*, preprint, arXiv:0806.4259.

[36] H. Matui, *Classification of outer actions of $\mathbb{Z}^N$ on $\mathcal{O}_2$*, Adv. Math. **217** (2008), 2872–2896.

[37] H. Matui, *$\mathbb{Z}$-actions on AH algebras and $\mathbb{Z}^2$-actions on AF algebras*, to appear in Comm. Math. Phys., arXiv:0907.2474.

[38] H. Matui and Y. Sato, *$\mathcal{Z}$-stability of crossed products by strongly outer actions*, preprint, arXiv:0912.4804.

[39] H. Nakamura, *The Rohlin property for $\mathbb{Z}^2$-actions on UHF algebras*, J. Math. Soc. Japan, **51** (1999), 583–612.

[40] H. Nakamura, *Aperiodic automorphisms of nuclear purely infinite simple C^*-algebras*, Ergodic Theory Dynam. Systems, **20** (2000), 1749–1765.

[41] M. Nakamura and Z. Takeda, *A Galois theory for finite factors.* Proc. Japan Acad. **36** (1960), 258–260.

[42] A. Ocneanu, *Actions of discrete amenable groups on von Neumann algebras*, Lecture Notes in Mathematics, 1138. Springer-Verlag, Berlin, 1985.

[43] H. Osaka and N. C. Phillips, *Furstenberg transformations on irrational rotation algebras*, Ergodic Theory Dynam. Systems, **26** (2006), 1623–1651.

[44] H. Osaka and N. C. Phillips, *Crossed products by finite group actions with the Rokhlin property*, to appear in Math. Z., arXiv:0704.3651.

[45] N. C. Phillips, *The tracial Rokhlin property for actions of finite groups on C^*-algebras*, preprint, arXiv:math/0609782.

[46] S. Popa and A. Wassermann, *Actions of compact Lie groups on von Neumann algebras*, C. R. Acad. Sci. Paris Ser. I Math. **315** (1992), 421–426.

[47] M. Rørdam, *Classification of nuclear, simple C^*-algebras*, Classification of nuclear C^*-algebras. Entropy in operator algebras, 1–145. Encyclopaedia Math. Sci., **126**, Springer, Berlin, 2002.

[48] Y. Sato, *The Rohlin property for automorphisms of the Jiang-Su algebra*, preprint, arXiv:0908.0135.

[49] N. Steenrod, *The topology of fibre bundles*, Princeton Landmarks in Mathematics. Princeton Paperbacks. Princeton University Press, Princeton, NJ, 1999.

[50] M. Takesaki, *Theory of operator algebras. I*, Encyclopaedia of Mathematical Sciences, 124. Operator Algebras and Non-commutative Geometry, 5. Springer-Verlag, Berlin, 2002.

[51] M. Takesaki, *Theory of operator algebras. II*, Encyclopaedia of Mathematical Sciences, 125. Operator Algebras and Non-commutative Geometry, 6. Springer-Verlag, Berlin, 2003.

[52] M. Takesaki, *Theory of operator algebras. III*, Encyclopaedia of Mathematical Sciences, 127. Operator Algebras and Non-commutative Geometry, 8. Springer-Verlag, Berlin, 2003.

[53] R. Tomatsu, *A characterization of right coideals of quotient type and its application to classification of Poisson boundaries*, Comm. Math. Phys. **275** (2007), 271–296.

[54] R. Tomatsu, *A Galois correspondence for compact quantum group actions*, J. Reine Angew. Math. **633** (2009), 165–182.

[55] S. Vaes and N. Vander Vennet, *Identification of the Poisson and Martin boundaries of orthogonal discrete quantum groups. J.* Inst. Math. Jussieu, **7** (2008), 391–412.

[56] S. Vaes and N. Vander Vennet, *Poisson boundary of the discrete quantum group* $\widehat{A_u(F)}$, to appear in Compositio Math., arXiv:0812.0804.

[57] W. Winter, *Localizing the Elliott conjecture at strongly self-absorbing C^*-algebras*, preprint, arXiv:0708.0283.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

L_1 Embeddings of the Heisenberg Group and Fast Estimation of Graph Isoperimetry

Assaf Naor*

Abstract

We survey connections between the theory of bi-Lipschitz embeddings and the Sparsest Cut Problem in combinatorial optimization. The story of the Sparsest Cut Problem is a striking example of the deep interplay between analysis, geometry, and probability on the one hand, and computational issues in discrete mathematics on the other. We explain how the key ideas evolved over the past 20 years, emphasizing the interactions with Banach space theory, geometric measure theory, and geometric group theory. As an important illustrative example, we shall examine recently established connections to the the structure of the Heisenberg group, and the incompatibility of its Carnot-Carathéodory geometry with the geometry of the Lebesgue space L_1.

Mathematics Subject Classification (2010). 46B85, 30L05, 46B80, 51F99.

Keywords. Bi-Lipschitz embeddings, Sparsest Cut Problem, Heisenberg group.

1. Introduction

Among the common definitions of the Heisenberg group $\mathbb{H}$, it will be convenient for us to work here with $\mathbb{H}$ modeled as $\mathbb{R}^3$, equipped with the group product $(a,b,c)\cdot(a',b',c') \stackrel{\text{def}}{=} (a+a', b+b', c+c'+ab'-ba')$. The integer lattice $\mathbb{Z}^3$ is then a discrete cocompact subgroup of $\mathbb{H}$, denoted by $\mathbb{H}(\mathbb{Z})$, which is generated by the finite symmetric set $\{(\pm1,0,0),(0,\pm1,0),(0,0,\pm1)\}$. The word metric on $\mathbb{H}(\mathbb{Z})$ induced by this generating set will be denoted by d_W.

As noted by Semmes [66], a differentiability result of Pansu [61] implies that the metric space $(\mathbb{H}(\mathbb{Z}), d_W)$ does not admit a bi-Lipschitz embedding

*Research supported in part by NSF grants CCF-0635078 and CCF-0832795, BSF grant 2006009, and the Packard Foundation.

New York University, Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA. E-mail: naor@cims.nyu.edu.

into $\mathbb{R}^n$ for any $n \in \mathbb{N}$. This was extended by Pauls [62] to bi-Lipschitz non-embeddability results of $(\mathbb{H}(\mathbb{Z}), d_W)$ into metric spaces with either lower or upper curvature bounds in the sense of Alexandrov. In [52, 27] it was observed that Pansu's differentiability argument extends to Banach space targets with the Radon-Nikodým property (see [14, Ch. 5]), and hence $\mathbb{H}(\mathbb{Z})$ does not admit a bi-Lipschitz embedding into, say, a Banach space which is either reflexive or is a separable dual; in particular $\mathbb{H}(\mathbb{Z})$ does not admit a bi-Lipschitz embedding into any $L_p(\mu)$ space, $1 < p < \infty$, or into the sequence space ℓ_1.

The embeddability of $\mathbb{H}(\mathbb{Z})$ into the function space $L_1(\mu)$, when μ is non-atomic, turned out to be much harder to settle. This question is of particular importance since it is well understood that for μ non-atomic, $L_1(\mu)$ is a space for which the differentiability results quoted above manifestly break down. Nevertheless, Cheeger and Kleiner [26, 25] introduced a novel notion of differentiability for which they could prove a differentiability theorem for Lipschitz maps from the Heisenberg group to $L_1(\mu)$, thus establishing that $\mathbb{H}(\mathbb{Z})$ does not admit a bi-Lipschitz embedding into any $L_1(\mu)$ space.

Another motivation for the $L_1(\mu)$ embeddability question for $\mathbb{H}(\mathbb{Z})$ originates from [52], where it was established that it is connected to the Sparsest Cut Problem in the field of combinatorial optimization. For this application it was of importance to obtain quantitative estimates in the $L_1(\mu)$ non-embeddability results for $\mathbb{H}(\mathbb{Z})$. It turns out that establishing such estimates is quite subtle, as they require overcoming finitary issues that do not arise in the infinite setting of [25, 28]. The following two theorems were proved in [29, 30]. Both theorems follow painlessly from a more general theorem that is stated and discussed in Section 5.4.

Theorem 1.1. *There exists a universal constant $c > 0$ such that any embedding into $L_1(\mu)$ of the restriction of the word metric d_W to the $n \times n \times n$ grid $\{1, \ldots, n\}^3$ incurs distortion $\gtrsim (\log n)^c$.*

Following Gromov [38], the compression rate of $f : \mathbb{H}(\mathbb{Z}) \to L_1(\mu)$, denoted $\omega_f(\cdot)$, is defined as the largest non-decreasing function such that for all $x, y \in \mathbb{H}(\mathbb{Z})$ we have $\|f(x) - f(y)\|_1 \geqslant \omega_f(d_W(x, y))$ (see [7] for more information on this topic).

Theorem 1.2. *There exists a universal constant $c > 0$ such that for every function $f : \mathbb{H}(\mathbb{Z}) \to L_1(\mu)$ which is 1-Lipschitz with respect to the word metric d_W, we have $\omega_f(t) \lesssim t/(\log t)^c$ for all $t \geqslant 2$.*

Evaluating the supremum of those $c > 0$ for which Theorem 1.1 holds true remains an important open question, with geometric significance as well as importance to theoretical computer science. Conceivably we could get c in Theorem 1.1 to be arbitrarily close to $\frac{1}{2}$, which would be sharp since the results of [8, 64] imply (see the explanation in [41]) that the metric space $\big(\{1, \ldots, n\}^3, d_W\big)$ embeds into ℓ_1 with distortion $\lesssim \sqrt{\log n}$. Similarly, we do not know the best possible c in Theorem 1.2; $\frac{1}{2}$ is again the limit here since it

was shown in [69] that there exists a 1-Lipschitz mapping $f : \mathbb{H}(\mathbb{Z}) \to \ell_1$ for which $\omega_f(t) \gtrsim t/(\sqrt{\log t} \cdot \log\log t)$.

The purpose of this article is to describe the above non-embeddability results for the Heisenberg group. Since one of the motivations for these investigations is the application to the Sparsest Cut Problem, we also include here a detailed discussion of this problem from theoretical computer science, and its deep connections to metric geometry. Our goal is to present the ideas in a way that is accessible to mathematicians who do not necessarily have background in computer science.

Acknowledgements. I am grateful to the following people for helpful comments and suggestions on earlier versions of this manuscript: Tim Austin, Keith Ball, Subhash Khot, Bruce Kleiner, Russ Lyons, Manor Mendel, Gideon Schechtman, Lior Silberman.

2. Embeddings

A metric space $(\mathscr{M}, d_{\mathscr{M}})$ is said to embed with distortion $D \geqslant 1$ into a metric space $(\mathscr{Y}, d_Y)$ if there exists a mapping $f : \mathscr{M} \to \mathscr{Y}$, and a scaling factor $s > 0$, such that for all $x, y \in \mathscr{M}$ we have $sd_{\mathscr{M}}(x, y) \leqslant d_{\mathscr{Y}}(f(x), f(y)) \leqslant Dsd_{\mathscr{M}}(x, y)$. The infimum over those $D \geqslant 1$ for which $(\mathscr{M}, d_{\mathscr{M}})$ embeds with distortion D into $(\mathscr{Y}, d_Y)$ is denoted by $c_{\mathscr{Y}}(\mathscr{M})$. If $(\mathscr{M}, d_{\mathscr{M}})$ does not admit a bi-Lipschitz embedding into $(\mathscr{Y}, d_Y)$, we will write $c_{\mathscr{Y}}(\mathscr{M}) = \infty$.

Throughout this paper, for $p \geqslant 1$, the space L_p will stand for $L_p([0, 1], \lambda)$, where λ is Lebesgue measure. The spaces ℓ_p and ℓ_p^n will stand for the space of p-summable infinite sequences, and $\mathbb{R}^n$ equipped with the ℓ_p norm, respectively. Much of this paper will deal with bi-Lipschitz embeddings of *finite* metric spaces into L_p. Since every n-point subset of an $L_p(\Omega, \mu)$ space embeds isometrically into $\ell_p^{n(n-1)/2}$ (see the discussion in [12]), when it comes to embeddings of finite metric spaces, the distinction between different $L_p(\Omega, \mu)$ spaces is irrelevant. Nevertheless, later, in the study of the embeddability of the Heisenberg group, we will need to distinguish between sequence spaces and function spaces.

For $p \geqslant 1$ we will use the shorter notation $c_p(\mathscr{M}) = c_{L_p}(\mathscr{M})$. The parameter $c_2(\mathscr{M})$ is known as the Euclidean distortion of $\mathscr{M}$. Dvoretzky's theorem says that if $\mathscr{Y}$ is an infinite dimensional Banach space then $c_{\mathscr{Y}}(\ell_2^n) = 1$ for all $n \in \mathbb{N}$. Thus, for every finite metric space $\mathscr{M}$ and every infinite dimensional Banach space $\mathscr{Y}$, we have $c_2(\mathscr{M}) \geqslant c_{\mathscr{Y}}(\mathscr{M})$.

The following famous theorem of Bourgain [15] will play a key role in what follows:

Theorem 2.1 (Bourgain's embedding theorem [15]). *For every n-point metric space $(\mathscr{M}, d_{\mathscr{M}})$, we have*

$$c_2(\mathscr{M}) \lesssim \log n. \tag{1}$$

Bourgain proved in [15] that the estimate (1) is sharp up to an iterated logarithm factor, i.e., that there exist arbitrarily large n-point metric spaces $\mathscr{M}_n$ for which $c_2(\mathscr{M}_n) \gtrsim \frac{\log n}{\log\log n}$. The $\log\log n$ term was removed in the important paper [56] of Linial, London and Rabinovich, who showed that the shortest path metric on bounded degree n-vertex expander graphs has Euclidean distortion $\gtrsim \log n$.

If one is interested only in embeddings into infinite dimensional Banach spaces, then Theorem 2.1 is stated in the strongest possible form: as noted above, it implies that for every infinite dimensional Banach space $\mathscr{Y}$, we have $c_{\mathscr{Y}}(\mathscr{M}) \lesssim \log n$. Below, we will actually use Theorem 2.1 for embeddings into L_1, i.e., we will use the fact that $c_1(\mathscr{M}) \lesssim \log n$. The expander based lower bound of Linial, London and Rabinovich [56] extends to embeddings into L_1 as well, i.e., even this weaker form of Bourgain's embedding theorem is asymptotically sharp. We refer to [58, Ch. 15] for a comprehensive discussion of these issues, as well as a nice presentation of the proof of Bourgain's embedding theorem.

3. L_1 as a Metric Space

Let (Ω,μ) be a measure space. Define a mapping $T : L_1(\Omega,\mu) \to L_\infty(\Omega\times\mathbb{R},\mu\times\lambda)$, where λ is Lebesgue measure, by:

$$T(f)(\omega,x) \stackrel{\text{def}}{=} \begin{cases} 1 & 0 < x \leqslant f(\omega), \\ -1 & f(\omega) < x < 0, \\ 0 & \text{otherwise.} \end{cases}$$

For all $f,g \in L_1(\Omega,\mu)$ we have:

$$\Big|T(f)(\omega,x) - T(g)(\omega,x)\Big| = \begin{cases} 1 & g(\omega) < x \leqslant f(\omega) \text{ or } f(\omega) < x \leqslant g(\omega), \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for all $p > 0$ we have,

$$\begin{aligned} \|T(f) - T(g)\|^p_{L_p(\Omega\times\mathbb{R},\mu\times\lambda)} &= \int_\Omega \left(\int_{(g(\omega),f(\omega)]\sqcup(f(\omega),g(\omega)]} d\lambda \right) d\mu(\omega) \\ &= \int_\Omega |f(\omega) - g(\omega)| d\mu(\omega) = \|f-g\|_{L_1(\Omega,\mu)}. \quad (2) \end{aligned}$$

Specializing (2) to $p = 2$, we see that:

$$\|T(f) - T(g)\|_{L_2(\Omega\times\mathbb{R},\mu\times\lambda)} = \sqrt{\|f-g\|_{L_1(\Omega,\mu)}}.$$

Corollary 3.1. *The metric space* $\left(L_1(\Omega,\mu), \|f-g\|^{1/2}_{L_1(\Omega,\mu)}\right)$ *admits an isometric embedding into Hilbert space.*

Another useful corollary is obtained when (2) is specialized to the case $p = 1$. Take an arbitrary finite subset $X \subseteq L_1(\Omega, \mu)$. For every $(\omega, x) \in \Omega \times \mathbb{R}$ consider the set $S(\omega, x) = \{f \in X : \ x \leqslant f(\omega)\} \subseteq X$. For every $S \subseteq X$ we can define a measurable subset $E_S = \{(\omega, x) \in \Omega \times \mathbb{R} : \ S(\omega, x) = S\} \subseteq \Omega \times \mathbb{R}$. By the definition of T, for every $f, g \in X$ we have

$$\begin{aligned} \|f - g\|_{L_1(\Omega,\mu)} &\overset{(2)}{=} \|T(f) - T(g)\|_{L_1(\Omega\times\mathbb{R},\mu\times\lambda)} \\ &= \int_{\Omega\times\mathbb{R}} \Big|\mathbf{1}_{S(w,x)}(f) - \mathbf{1}_{S(w,x)}(g)\Big| \, d(\mu\times\lambda)(\omega, x) \\ &= \sum_{S\subseteq X} (\mu\times\lambda)(E_S)\Big|\mathbf{1}_S(f) - \mathbf{1}_S(g)\Big|, \end{aligned}$$

where here, and in what follows, $\mathbf{1}_S(\cdot)$ is the characteristic function of S. Writing $\beta_S = (\mu \times \lambda)(E_S)$, we have the following important corollary:

Corollary 3.2. *Let $X \subseteq L_1(\Omega, \mu)$ be a finite subset of $L_1(\Omega, \mu)$. Then there exist nonnegative numbers $\{\beta_S\}_{S\subseteq X} \subseteq [0, \infty)$ such that for all $f, g \in X$ we have:*

$$\|f - g\|_{L_1(\Omega,\mu)} = \sum_{S\subseteq X} \beta_S\Big|\mathbf{1}_S(f) - \mathbf{1}_S(g)\Big|. \tag{3}$$

A metric space $(\mathscr{M}, d_{\mathscr{M}})$ is said to be of *negative type* if the metric space $\left(\mathscr{M}, d_{\mathscr{M}}^{1/2}\right)$ admits an isometric embedding into Hilbert space. Such metrics will play a crucial role in the ensuing discussion. This terminology (see e.g., [33]) is due to a classical theorem of Schoenberg [65], which asserts that $(\mathscr{M}, d_{\mathscr{M}})$ is of negative type if and only if for every $n \in \mathbb{N}$ and every $x_1, \ldots, x_n \in X$, the matrix $(d_{\mathscr{M}}(x_i, x_j))_{i,j=1}^n$ is negative semidefinite on the orthogonal complement of the main diagonal in $\mathbb{C}^n$, i.e., for all $\zeta_1, \ldots, \zeta_n \in \mathbb{C}$ with $\sum_{j=1}^n \zeta_j = 0$ we have $\sum_{i=1}^n \sum_{j=1}^n \zeta_i \overline{\zeta_j} d_{\mathscr{M}}(x_i, x_j) \leqslant 0$. Corollary (3.1) can be restated as saying that $L_1(\Omega, \mu)$ is a metric space of negative type.

Corollary (3.2) is often called the *cut cone representation of L_1 metrics*. To explain this terminology, consider the set $\mathscr{C} \subseteq \mathbb{R}^{n^2}$ of all $n \times n$ real matrices $A = (a_{ij})$ such that there is a measure space (Ω, μ) and $f_1, \ldots, f_n \in L_1(\Omega, \mu)$ with $a_{ij} = \|f_i - f_j\|_{L_1(\Omega,\mu)}$ for all $i, j \in \{1, \ldots, n\}$. If $f_1, \ldots, f_n \in L_1(\Omega_1, \mu_1)$ and $g_1, \ldots, g_n \in L_1(\Omega_2, \mu_2)$ then for all $c_1, c_2 \geqslant 0$ and $i, j \in \{1, \ldots, n\}$ we have

$$c_1\|f_i - f_j\|_{L_1(\Omega_1,\mu_1)} + c_2\|f_i - f_j\|_{L_1(\Omega_2,\mu_2)} = \|h_i - h_j\|_{L_1(\Omega_1\sqcup\Omega_2,\mu_1\sqcup\mu_2)},$$

where $h_1, \ldots, h_n$ are functions defined on the disjoint union $\Omega_1 \sqcup \Omega_2$ as follows: $h_i(\omega) = c_1 f_i(\omega)\mathbf{1}_{\Omega_1}(\omega) + c_2 g_i(\omega)\mathbf{1}_{\Omega_2}(\omega)$. This observation shows that $\mathscr{C}$ is a cone (of dimension $n(n-1)/2$). Identity (3) says that the cone $\mathscr{C}$ is generated by the rays induced by cut semimetrics, i.e., by matrices of the form $a_{ij} = |\mathbf{1}_S(i) - \mathbf{1}_S(j)|$ for some $S \subseteq \{1, \ldots, n\}$. It is not difficult to see that these rays are actually the extreme rays of the cone $\mathscr{C}$. Carathéodory's theorem (for cones) says that we can choose the coefficients $\{\beta_S\}_{S\subseteq X}$ in (3) so that only $n(n-1)/2$ of them are non-zero.

4. The Sparsest Cut Problem

Given $n \in \mathbb{N}$ and two symmetric functions $C, D : \{1, \ldots, n\} \times \{1, \ldots, n\} \to [0, \infty)$ (called capacities and demands, respectively), and a subset $\emptyset \neq S \subsetneq \{1, \ldots, n\}$, write

$$\Phi(S) \stackrel{\mathrm{def}}{=} \frac{\sum_{i=1}^n \sum_{j=1}^n C(i,j) \cdot |\mathbf{1}_S(i) - \mathbf{1}_S(j)|}{\sum_{i=1}^n \sum_{j=1}^n D(i,j) \cdot |\mathbf{1}_S(i) - \mathbf{1}_S(j)|}. \tag{4}$$

The value

$$\Phi^*(C, D) \stackrel{\mathrm{def}}{=} \min_{\emptyset \neq S \subsetneq \{1,\ldots,n\}} \Phi(S) \tag{5}$$

is the minimum over all cuts (two-part partitions) of $\{1, \ldots, n\}$ of the ratio between the total capacity crossing the boundary of the cut and the total demand crossing the boundary of the cut.

Finding in polynomial time a cut for which $\Phi^*(C, D)$ is attained up to a definite multiplicative constant is called the Sparsest Cut problem. This problem is used as a subroutine in many approximation algorithms for NP-hard problems; see the survey articles [68, 22], as well as [53, 1] and the references in [6, 5] for some of the vast literature on this topic. Computing $\Phi^*(C, D)$ exactly has been long known to be NP-hard [67]. More recently, it was shown in [31] that there exists $\varepsilon_0 > 0$ such that it is NP-hard to approximate $\Phi^*(C, D)$ to within a factor smaller than $1 + \varepsilon_0$. In [47, 24] it was shown that it is Unique Games hard to approximate $\Phi^*(C, D)$ to within any constant factor (see [44, 45] for more information on the Unique Games Conjecture; we will return to this issue in Section 4.3.3).

It is customary in the literature to highlight the support of the capacities function C: this allows us to introduce a particulary important special case of the Sparsest Cut Problem. Thus, a different way to formulate the above setup is via an n-vertex graph $G = (V, E)$, with a positive weight (called a capacity) $C(e)$ associated to each edge $e \in E$, and a nonnegative weight (called a demand) $D(u, v)$ associated to each pair of vertices $u, v \in V$. The goal is to evaluate in polynomial time (and in particular, while examining only a negligible fraction of the subsets of V) the quantity:

$$\Phi^*(C, D) = \min_{\emptyset \neq S \subsetneq V} \frac{\sum_{uv \in E} C(uv) \, |\mathbf{1}_S(u) - \mathbf{1}_S(v)|}{\sum_{u,v \in V} D(u,v) \, |\mathbf{1}_S(u) - \mathbf{1}_S(v)|}.$$

To get a feeling for the meaning of Φ^*, consider the case $C(e) = D(u, v) = 1$ for all $e \in E$ and $u, v \in V$. This is an important instance of the Sparsest Cut problem which is called "Sparsest Cut with Uniform Demands". In this case Φ^* becomes:

$$\Phi^* = \min_{\emptyset \neq S \subsetneq V} \frac{\#\{\text{edges joining } S \text{ and } V \setminus S\}}{|S| \cdot |V \setminus S|}.$$

Thus, in the case of uniform demands, the Sparsest Cut problem essentially amounts to solving efficiently the combinatorial isoperimetric problem on G: determining the subset of the graph whose ratio of edge boundary to its size is as small as possible.

In the literature it is also customary to emphasize the size of the support of the demand function D, i.e., to state bounds in terms of the number k of pairs $\{i,j\} \subseteq \{1,\ldots,n\}$ for which $D(i,j) > 0$. For the sake of simplicity of exposition, we will not adopt this convention here, and state all of our bounds in terms of n rather than the number of positive demand pairs k. We refer to the relevant references for the simple modifications that are required to obtain bounds in terms of k alone.

From now on, the Sparsest Cut problem will be understood to be with general capacities and demands; when discussing the special case of uniform demands we will say so explicitly. In applications, general capacities and demands are used to tune the notion of "interface" between S and $V \setminus S$ to a wide variety of combinatorial optimization problems, which is one of the reasons why the Sparsest Cut problem is so versatile in the field of approximation algorithms.

4.1. Reformulation as an optimization problem over L_1.

Although the Sparsest Cut Problem clearly has geometric flavor as a discrete isoperimetric problem, the following key reformulation of it, due to [11, 56], explicitly relates it to the geometry of L_1.

Lemma 4.1. *Given symmetric $C, D : \{1,\ldots,n\} \times \{1,\ldots,n\} \to [0,\infty)$, we have:*

$$\Phi^*(C,D) = \min_{f_1,\ldots,f_n \in L_1} \frac{\sum_{i=1}^n \sum_{j=1}^n C(i,j)\|f_i - f_j\|_1}{\sum_{i=1}^n \sum_{j=1}^n D(i,j)\|f_i - f_j\|_1}. \tag{6}$$

Proof. Let ϕ denote the right hand side of (6), and write $\Phi^* = \Phi^*(C,D)$. Given a subset $S \subseteq \{1,\ldots,n\}$, by considering $f_i = \mathbf{1}_S(i) \in \{0,1\} \subseteq L_1$ we see that that $\phi \leqslant \Phi^*$. In the reverse direction, if $X = \{f_1,\ldots,f_n\} \subseteq L_1$ then let $\{\beta_S\}_{S \subseteq X}$ be the non-negative weights from Corollary 3.2. For $S \subseteq X$ define a subset of $\{1,\ldots,n\}$ by $S' = \{i \in \{1,\ldots,n\} : \ f_i \in S\}$. It follows from the definition of Φ^* that for all $S \subseteq X$ we have,

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^n C(i,j)|\mathbf{1}_S(f_i) - \mathbf{1}_S(f_j)| &\overset{(4)}{=} \Phi(S') \sum_{i=1}^n \sum_{j=1}^n D(i,j)|\mathbf{1}_S(f_i) - \mathbf{1}_S(f_j)| \\ &\overset{(5)}{\geqslant} \Phi^* \sum_{i=1}^n \sum_{j=1}^n D(i,j)|\mathbf{1}_S(f_i) - \mathbf{1}_S(f_j)|. \end{aligned} \tag{7}$$

Thus

$$\sum_{i=1}^n\sum_{j=1}^n C(i,j)\|f_i-f_j\|_1 \overset{(3)}{=} \sum_{S\subseteq X}\beta_S\sum_{i=1}^n\sum_{j=1}^n C(i,j)|\mathbf{1}_S(f_i)-\mathbf{1}_S(f_j)|$$
$$\overset{(7)}{\geqslant} \Phi^*\sum_{S\subseteq X}\beta_S\sum_{i=1}^n\sum_{j=1}^n D(i,j)|\mathbf{1}_S(f_i)-\mathbf{1}_S(f_j)| \overset{(3)}{=} \sum_{i=1}^n\sum_{j=1}^n D(i,j)\|f_i-f_j\|_1.$$

It follows that $\phi\geqslant\Phi^*$, as required. □

4.2. The linear program. Lemma 4.1 is a reformulation of the Sparsest Cut Problems in terms of a continuous optimization problem on the space L_1. Being a reformulation, it shows in particular that solving L_1 optimization problems such as the right hand side of (6) is NP-hard.

In the beautiful paper [53] of Leighton and Rao it was shown that there exists a polynomial time algorithm that, given an n-vertex graph $G=(V,E)$, computes a number which is guaranteed to be within a factor of $\lesssim\log n$ of the uniform Sparsest Cut value (4). The Leighton-Rao algorithm uses combinatorial ideas which do not apply to Sparsest Cut with general demands. A breakthrough result, due to Linial-London-Rabinovich [56] and Aumann-Rabani [9], introduced embedding methods to this field, yielding a polynomial time algorithm which computes $\Phi^*(C,D)$ up to a factor $\lesssim\log n$ for all $C,D:\{1,\ldots,n\}\times\{1,\ldots,n\}\to[0,\infty)$.

The key idea of [56, 9] is based on replacing the finite subset $\{f_1,\ldots,f_n\}$ of L_1 in (6) by an *arbitrary* semimetric on $\{1,\ldots,n\}$. Specifically, by homogeneity we can always assume that the denominator in (6) equals 1, in which case Lemma 4.1 says that $\Phi^*(C,D)$ equals the minimum of $\sum_{i=1}^n\sum_{j=1}^n C(i,j)d_{ij}$, given that $\sum_{i=1}^n\sum_{j=1}^n D(i,j)d_{ij}=1$ and there exist $f_1,\ldots,f_n\in L_1$ for which $d_{ij}=\|f_i-f_j\|_1$ for all $i,j\in\{1,\ldots,n\}$. We can now ignore the fact that d_{ij} was a semimetric that came from a subset of L_1, i.e., we can define $M^*(C,D)$ to be the minimum of $\sum_{i=1}^n\sum_{j=1}^n C(i,j)d_{ij}$, given that $\sum_{i=1}^n\sum_{j=1}^n D(i,j)d_{ij}=1$, $d_{ii}=0$, $d_{ij}\geqslant 0$, $d_{ij}=d_{ji}$ for all $i,j\in\{1,\ldots,n\}$ ($n(n-1)/2$ symmetry constraints) and $d_{ij}\leqslant d_{ik}+d_{kj}$ for all $i,j,k\in\{1,\ldots,n\}$ ($\leqslant n^3$ triangle inequality constraints).

Clearly $M^*(C,D)\leqslant\Phi^*(C,D)$, since we are minimizing over all semimetrics rather than just those arising from subsets of L_1. Moreover, $M^*(C,D)$ can be computed in polynomial time up to arbitrarily good precision [40], since it is a linear program (minimizing a linear functional in the variables (d_{ij}) subject to polynomially many linear constraints).

The linear program produces a semimetric d^*_{ij} on $\{1,\ldots,n\}$ which satisfies $M^*(C,D)=\sum_{i=1}^n\sum_{j=1}^n C(i,j)d^*_{ij}$ and $\sum_{i=1}^n\sum_{j=1}^n D(i,j)d^*_{ij}=1$ (ignoring arbitrarily small errors). By Lemma 4.1 we need to somehow relate this semimetric to L_1. It is at this juncture that we see the power of Bourgain's embedding theorem 2.1: the constraints of the linear program only provide us the information

that d^*_{ij} is a semimetric, and nothing else. So, we need to be able to somehow handle arbitrary metric spaces—precisely what Bourgain's theorem does, by furnishing $f_1, \ldots, f_n \in L_1$ such that for all $i, j \in \{1, \ldots, n\}$ we have

$$\frac{d^*_{ij}}{\log n} \lesssim \|f_i - f_j\|_1 \leqslant d^*_{ij}. \tag{8}$$

Now,

$$\begin{aligned}\Phi^*(C,D) &\overset{(6)}{\leqslant} \frac{\sum_{i=1}^n \sum_{j=1}^n C(i,j)\|f_i - f_j\|_1}{\sum_{i=1}^n \sum_{j=1}^n D(i,j)\|f_i - f_j\|_1} \\ &\overset{(8)}{\lesssim} \log n \cdot \frac{\sum_{i=1}^n \sum_{j=1}^n C(i,j) d^*_{ij}}{\sum_{i=1}^n \sum_{j=1}^n D(i,j) d^*_{ij}} = \log n \cdot M^*(C,D). \end{aligned} \tag{9}$$

Thus, $\frac{\Phi^*(C,D)}{\log n} \lesssim M^*(C,D) \leqslant \Phi^*(C,D)$, i.e., the polynomial time algorithm of computing $M^*(C,D)$ is guranteed to produce a number which is within a factor $\lesssim \log n$ of $\Phi^*(C,D)$.

Remark 4.2. In the above argument we only discussed the algorithmic task of fast estimation of the number $\Phi^*(C,D)$, rather than the problem of producing in polynomial time a subset $\emptyset \neq S \subsetneq \{1, \ldots, n\}$ for which $\Phi^*(S)$ is close up to a certain multiplicative guarantee to the optimum value $\Phi^*(C,D)$. All the algorithms discussed in this paper produce such a set S, rather than just approximating the number $\Phi^*(C,D)$. In order to modify the argument above to this setting, one needs to go into the proof of Bourgain's embedding theorem, which as currently stated as just an existential result for $f_1, \ldots, f_n$ as in (8). This issue is addressed in [56], which provides an algorithmic version of Bourgain's theorem. Ensuing algorithms in this paper can be similarly modified to produce a good cut S, but we will ignore this issue from now on, and continue to focus solely on algorithms for approximate computation of $\Phi^*(C,D)$.

4.3. The semidefinite program.

We have already stated in Section 2 that the logarithmic loss in the application (8) of Bourgain's theorem cannot be improved. Thus, in order to obtain a polynomial time algorithm with approximation guarantee better than $\lesssim \log n$, we need to impose additional geometric restrictions on the metric d^*_{ij}; conditions that will hopefully yield a class of metric spaces for which one can prove an L_1 distortion bound that is asymptotically smaller than the $\lesssim \log n$ of Bourgain's embedding theorem. This is indeed possible, based on a quadratic variant of the discussion in Section 4.2; an approach due to Goemans and Linial [37, 55, 54].

The idea of Goemans and Linial is based on Corollary 3.1, i.e., on the fact that the metric space L_1 is of negative type. We define $M^{**}(C,D)$ to be the minimum of $\sum_{i=1}^n \sum_{j=1}^n C(i,j) d_{ij}$, subject to the constraint that $\sum_{i=1}^n \sum_{j=1}^n D(i,j) d_{ij} = 1$ and d_{ij} is a semimetric of negative type on $\{1, \ldots, n\}$.

The latter condition can be equivalently restated as the requirement that, in addition to d_{ij} being a semimetric on $\{1,\ldots,n\}$, there exist vectors $v_1,\ldots,v_n \in L_2$ such that $d_{ij} = \|v_i - v_j\|_2^2$ for all $i,j \in \{1,\ldots,n\}$. Equivalently, there exists a symmetric positive semidefinite $n \times n$ matrix (a_{ij}) (the Gram matrix of $v_1,\ldots,v_n$), such that $d_{ij} = a_{ii} + a_{jj} - 2a_{ij}$ for all $i,j \in \{1,\ldots,n\}$.

Thus, $M^{**}(C,D)$ is the minimum of $\sum_{i=1}^n \sum_{j=1}^n C(i,j)(a_{ii} + a_{jj} - 2a_{ij})$, a linear function in the variables (a_{ij}), subject to the constraint that (a_{ij}) is a symmetric positive semidefinite matrix, in conjunction with the linear constraints $\sum_{i=1}^n \sum_{j=1}^n D(i,j)(a_{ii} + a_{jj} - 2a_{ij}) = 1$ and for all $i,j,k \in \{1,\ldots,n\}$, the triangle inequality constraint $a_{ii} + a_{jj} - 2a_{ij} \leqslant (a_{ii} + a_{kk} - 2a_{ik}) + (a_{kk} + a_{jj} - 2a_{kj})$. Such an optimization problem is called a semidefinite program, and by the methods described in [40], $M^{**}(C,D)$ can be computed with arbitrarily good precision in polynomial time.

Corollary 3.1 and Lemma 4.1 imply that $M^*(C,D) \leqslant M^{**}(C,D) \leqslant \Phi^*(C,D)$. The following breakthrough result of Arora, Rao and Vazirani [6] shows that for Sparsest Cut with uniform demands the Goemans-Linial approach does indeed yield an improved approximation algorithm:

Theorem 4.3 ([6]). *In the case of uniform demands, i.e., if $C(i,j) \in \{0,1\}$ and $D(i,j) = 1$ for all $i,j \in \{1,\ldots,n\}$, we have*

$$\frac{\Phi^*(C,D)}{\sqrt{\log n}} \lesssim M^{**}(C,D) \leqslant \Phi^*(C,D). \tag{10}$$

In the case of general demands we have almost the same result, up to lower order factors:

Theorem 4.4 ([5]). *For all symmetric $C,D : \{1,\ldots,n\} \times \{1,\ldots,n\} \to [0,\infty)$ we have*

$$\frac{\Phi^*(C,D)}{(\log n)^{\frac{1}{2}+o(1)}} \lesssim M^{**}(C,D) \leqslant \Phi^*(C,D). \tag{11}$$

The $o(1)$ term in (11) is $\lesssim \frac{\log\log\log n}{\log\log n}$. We conjecture that it could be removed altogether, though at present it seems to be an inherent artifact of complications in the proof in [5].

Before explaining some of the ideas behind the proofs of Theorem 4.3 and Theorem 4.4 (the full details are quite lengthy and are beyond the scope of this survey), we prove, following [58, Prop. 15.5.2], a crucial identity (attributed in [58] to Y. Rabinovich) which reformulates these results in terms of an L_1 embeddability problem.

Lemma 4.5. *We have*

$$\sup\left\{\frac{\Phi^*(C,D)}{M^{**}(C,D)} : C,D : \{1,\ldots,n\} \times \{1,\ldots,n\} \to (0,\infty)\right\}$$
$$= \sup\Big\{c_1\big(\{1,\ldots,n\},d\big) : d \text{ is a metric of negative type}\Big\}. \tag{12}$$

Proof. The proof of the fact that the left hand side of (12) is at most the right hand side of (12) is identical to the way (9) was deduced from (8).

In the reverse direction, let d^* be a metric of negative type on $\{1,\ldots,n\}$ for which $c_1(\{1,\ldots,n\},d^*) \stackrel{\text{def}}{=} c$ is maximal among all such metrics. Let $\mathscr{C} \subseteq \mathbb{R}^{n^2}$ be the cone in the space of $n \times n$ symmetric matrices from the last paragraph of Section 3, i.e., $\mathscr{C}$ consists of all matrices of the form $(\|f_i - f_j\|_1)$ for some $f_1,\ldots,f_n \in L_1$.

Fix $\varepsilon \in (0, c-1)$ and let $\mathscr{K}_\varepsilon \subseteq \mathbb{R}^{n^2}$ be the set of all symmetric matrices (a_{ij}) for which there exists $s > 0$ such that $sd^*(i,j) \leqslant a_{ij} \leqslant (c-\varepsilon)sd^*(i,j)$ for all $i,j \in \{1,\ldots,n\}$. By the definition of c, the convex sets $\mathscr{C}$ and $\mathscr{K}_\varepsilon$ are disjoint, since otherwise d^* would admit an embedding into L_1 with distortion $c-\varepsilon$. It follows that there exists a symmetric matrix $(h^\varepsilon_{ij}) \in \mathbb{R}^{n^2} \setminus \{0\}$ and $\alpha \in \mathbb{R}$, such that $\sum_{i=1}^n \sum_{j=1}^n h^\varepsilon_{ij} a_{ij} \leqslant \alpha$ for all $(a_{ij}) \in \mathscr{K}_\varepsilon$, and $\sum_{i=1}^n \sum_{j=1}^n h^\varepsilon_{ij} b_{ij} \geqslant \alpha$ for all $(b_{ij}) \in \mathscr{C}$. Since both $\mathscr{C}$ and $\mathscr{K}_\varepsilon$ are closed under multiplication by positive scalars, necessarily $\alpha = 0$.

Define $C^\varepsilon(i,j) \stackrel{\text{def}}{=} h^\varepsilon_{ij} \mathbf{1}_{\{h^\varepsilon_{ij} \geqslant 0\}}$ and $D^\varepsilon(i,j) \stackrel{\text{def}}{=} |h^\varepsilon_{ij}| \mathbf{1}_{\{h^\varepsilon_{ij} \leqslant 0\}}$. By definition of $M^{**}(C^\varepsilon, D^\varepsilon)$,

$$\sum_{i=1}^n \sum_{j=1}^n C^\varepsilon(i,j) d^*_{ij} \geqslant M^{**}(C^\varepsilon, D^\varepsilon) \cdot \sum_{i=1}^n \sum_{j=1}^n D^\varepsilon(i,j) d^*_{ij}. \tag{13}$$

By considering $a_{ij} \stackrel{\text{def}}{=} \left((c-\varepsilon)\mathbf{1}_{\{h^\varepsilon_{ij} \geqslant 0\}} + \mathbf{1}_{\{h^\varepsilon_{ij} < 0\}}\right) d^*(i,j) \in \mathscr{K}_\varepsilon$, the inequality $\sum_{i=1}^n \sum_{j=1}^n h^\varepsilon_{ij} a_{ij} \leqslant 0$ becomes:

$$\sum_{i=1}^n \sum_{j=1}^n D^\varepsilon(i,j) d^*_{ij} \geqslant (c-\varepsilon) \sum_{i=1}^n \sum_{j=1}^n C^\varepsilon(i,j) d^*_{ij}. \tag{14}$$

A combination of (13) and (14) implies that $(c-\varepsilon) M^{**}(C^\varepsilon, D^\varepsilon) \leqslant 1$. At the same time, for all $f_1,\ldots,f_n \in L_1$, the inequality $\sum_{i=1}^n \sum_{j=1}^n h^\varepsilon_{ij} \|f_i - f_j\|_1 \geqslant 0$ is the same as $\sum_{i=1}^n \sum_{j=1}^n C^\varepsilon(i,j) \|f_i - f_j\|_1 \geqslant \sum_{i=1}^n \sum_{j=1}^n D^\varepsilon(i,j) \|f_i - f_j\|_1$, which by Lemma 6 means that $\Phi^*(C^\varepsilon, D^\varepsilon) \geqslant 1$. Thus $\Phi^*(C^\varepsilon, D^\varepsilon)/M^{**}(C^\varepsilon, D^\varepsilon) \geqslant c-\varepsilon$, and since this holds for all $\varepsilon \in (0, c-1)$, the proof of Lemma 4.5 is complete. □

In the case of Sparsest Cut with uniform demands, we have the following result which is analogous to Lemma 4.5, where the L_1 bi-Lipschitz distortion is replaced by the smallest possible factor by which 1-Lipschitz functions into L_1 can distort the *average distance*. The proof is a slight variant of the proof of Lemma 4.5; the simple details are left to the reader. This connection between Sparsest Cut with uniform demands and embeddings that preserve the average distance is due to Rabinovich [63].

Lemma 4.6. *The supremum of $\Phi^*(C,D)/M^{**}(C,D)$ over all instances of uniform demands, i.e., when $C(i,j) \in \{0,1\}$ and $D(i,j) = 1$ for all $i,j \in$*

$\{1,\ldots,n\}$, *equals the infimum over* $A > 0$ *such that for all metrics* d *on* $\{1,\ldots,n\}$ *of negative type, there exist* $f_1,\ldots,f_n \in L_1$ *satisfying* $\|f_i - f_j\|_1 \leqslant d(i,j)$ *for all* $i,j \in \{1,\ldots,n\}$ *and* $A\sum_{i=1}^n\sum_{j=1}^n \|f_i - f_j\|_1 \geqslant \sum_{i=1}^n\sum_{j=1}^n d(i,j)$.

4.3.1. L_2 embeddings of negative type metrics. The proof of Theorem 4.3 in [6] is based on a clever geometric partitioning procedure for metrics of negative type. Building heavily on ideas of [6], in conjunction with some substantial additional combinatorial arguments, an alternative approach to Theorem 4.3 was obtained in [59], based on a purely graph theoretical statement which is of independent interest. We shall now sketch this approach, since it is modular and general, and as such it is useful for additional geometric corollaries. We refer to [59] for more information on these additional applications, as well as to [6] for the original proof of Theorem 4.3.

Let $G = (V,E)$ be an n-vertex graph. The *vertex expansion* of G, denoted $h(G)$, is the largest $h \geqslant 0$ such that every $S \subseteq V$ with $|S| \leqslant n/2$ has at least $h|S|$ neighbors in $V \setminus S$. The *edge expansion* of G, denoted $\alpha(G)$, is the largest $\alpha \geqslant 0$ such that for every $S \subseteq V$ with $|S| \leqslant n/2$, the number of edges joining S and $V \setminus S$ is at least $\alpha|S| \cdot \frac{|E|}{n}$. The main combinatorial statement of [59] relates these two notions of expansion of graphs:

Theorem 4.7 (Edge Replacement Theorem [59]). *For every graph* $G = (V,E)$ *with* $h(G) \geqslant \frac{1}{2}$ *there is a set of edges* E' *on* V *with* $\alpha(V,E') \gtrsim 1$*, and such that for every* $uv \in E'$ *we have* $d_G(u,v) \lesssim \sqrt{\log|V|}$*. Here* d_G *is the shortest path metric on* G *(with respect to the original edge set* E*), and all implicit constants are universal.*

It is shown in [59] that the $\lesssim \sqrt{\log n}$ bound on the length of the new edges in Theorem 4.7 is asymptotically tight. The proof of Theorem 4.7 is involved, and cannot be described here: it has two components, a combinatorial construction, as well a purely Hilbertian geometric argument based on, and simpler than, the original algorithm of [6]. We shall now explain how Theorem 4.7 implies Theorem 4.3 (this is somewhat different from the deduction in [59], which deals with a different semidefinite program for Sparsest Cut with uniform demands).

Proof of Theorem 4.3 assuming Theorem 4.7. An application of (the easy direction of) Lemma 4.6 shows that in order to prove Theorem 4.3 it suffices to show that if $(\mathscr{M},d)$ is an n-point metric space of negative type, with $\frac{1}{n^2}\sum_{x,y\in\mathscr{M}} d(x,y) = 1$, then there exists a mapping $F : \mathscr{M} \to \mathbb{R}$ which is 1-Lipschitz and such that $\frac{1}{n^2}\sum_{x,y\in\mathscr{M}} |F(x) - F(y)| \gtrsim 1/\sqrt{\log n}$. In what follows we use the standard notation for closed balls: for $x \in \mathscr{M}$ and $t \geqslant 0$, set $B(x,t) = \{y \in \mathscr{M} :\ d(x,y) \leqslant t\}$.

Choose $x_0 \in \mathscr{M}$ with $\frac{1}{n}\sum_{y\in\mathscr{M}} d(x_0,y) = r \stackrel{\text{def}}{=} \min_{x\in\mathscr{M}} \frac{1}{n}\sum_{y\in\mathscr{M}} d(x,y)$. Then $r \leqslant \frac{1}{n^2}\sum_{x,y\in\mathscr{M}} d(x,y) = 1$, implying $1 \geqslant \frac{1}{n}\sum_{y\in\mathscr{M}} d(x_0,y) > \frac{2}{n}|\mathscr{M} \setminus B(x_0,2)|$, or $|B(x_0,2)| > n/2$. Similarly $|B(x_0,4)| > 3n/4$.

Assume first that $\frac{1}{n^2}\sum_{x,y\in B(x_0,4)} d(x,y) \leqslant \frac{1}{4}$ (this will be the easy case). Then

$$1 = \frac{1}{n^2}\sum_{x,y\in\mathscr{M}} d(x,y) \leqslant \frac{1}{4} + \frac{2}{n^2}\sum_{x\in\mathscr{M}}\sum_{y\in\mathscr{M}\setminus B(x_0,4)}\Big(d(x,x_0)+d(x_0,y)\Big)$$
$$= \frac{1}{4}+\frac{2r}{n}|\mathscr{M}\setminus B(x_0,4)|+\frac{2}{n}\sum_{y\in\mathscr{M}\setminus B(x_0,4)} d(x_0,y) \leqslant \frac{3}{4}+\frac{2}{n}\sum_{y\in\mathscr{M}\setminus B(x_0,4)} d(x_0,y),$$

or $\frac{1}{n}\sum_{y\in\mathscr{M}\setminus B(x_0,4)} d(x_0,y) \geqslant \frac{1}{8}$. Define a 1-Lipschitz mapping $F:\mathscr{M}\to\mathbb{R}$ by $F(x) = d\big(x, B(x_0,2)\big) = \min_{y\in B(x_0,2)} d(x,y)$. The triangle inequality implies that for every $y\in\mathscr{M}\setminus B(x_0,4)$ we have $F(y)\geqslant \frac{1}{2}d(y,x_0)$. Thus

$$\frac{1}{n^2}\sum_{x,y\in\mathscr{M}}|F(x)-F(y)| \geqslant \frac{|B(x_0,2)|}{n^2}\sum_{y\in\mathscr{M}\setminus B(x_0,4)} d\big(y,B(x_0,2)\big)$$
$$> \frac{1}{2n}\sum_{y\in\mathscr{M}\setminus B(x_0,4)}\frac{1}{2}d(y,x_0) \gtrsim 1 = \frac{1}{n^2}\sum_{x,y\in\mathscr{M}} d(x,y).$$

This completes the easy case, where there is even no loss of $1/\sqrt{\log n}$ (and we did not use yet the assumption that d is a metric of negative type).

We may therefore assume from now on that $\frac{1}{n^2}\sum_{x,y\in B(x_0,4)} d(x,y) \geqslant \frac{1}{4}$. The fact that d is of negative type means that there are vectors $\{v_x\}_{x\in\mathscr{M}}\subseteq L_2$ such that $d(x,y)=\|v_x-v_y\|_2^2$ for all $x,y\in\mathscr{M}$.

We will show that for a small enough universal constant $\varepsilon>0$, there are two sets $S_1,S_2\subseteq B(x_0,4)$ such that $|S_1|,|S_2|\geqslant \varepsilon n$ and $d(S_1,S_2)\geqslant \varepsilon^2/\sqrt{\log n}$. Once this is achieved, the mapping $F:\mathscr{M}\to\mathbb{R}$ given by $F(x)=d(x,S_1)$ will satisfy $\frac{1}{n^2}\sum_{x,y\in\mathscr{M}}|F(x)-F(y)| \geqslant \frac{2}{n^2}|S_1|\cdot|S_2|\frac{\varepsilon^2}{\sqrt{\log n}} \geqslant \frac{2\varepsilon^4}{\sqrt{\log n}}$, as desired.

Assume for contradiction that no such S_1,S_2 exist. Define a set of edges E_0 on $B(x_0,4)$ by $E_0 \stackrel{\text{def}}{=} \Big\{\{x,y\}\subseteq B(x_0,4):\ x\neq y\ \wedge\ d(x,y)<\varepsilon^2/\sqrt{\log n}\Big\}$. Our contrapositive assumption says that any two subsets $S_1,S_2\subseteq B(x_0,4)$ with $|S_1|,|S_2|\geqslant\varepsilon n\geqslant\varepsilon|B(x_0,4)|$ are joined by an edge from E_0. By a (simple) general graph theoretical lemma (see [59, Lem 2.3]), this implies that, provided $\varepsilon\leqslant 1/10$, there exists a subset $V\subseteq B(x_0,4)$ with $|V|\geqslant(1-\varepsilon)|B(x_0,4)|\gtrsim n$, such that the graph induced by E_0 on V, i.e., $G=\Big(V, E=E_0\cap\binom{V}{2}\Big)$, has $h(G)\geqslant\frac{1}{2}$.

We are now in position to apply the Edge Replacement Theorem, i.e., Theorem 4.7. We obtain a new set of edges E' on V such that $\alpha(V,E')\gtrsim 1$ and for every $xy\in E'$ we have $d_G(x,y)\lesssim\sqrt{\log n}$. The latter condition means that there exists a path $\{x=x_0,x_1,\ldots,x_m=y\}\subseteq V$ such that $m\lesssim\sqrt{\log n}$ and $x_ix_{i-1}\in E$ for every $i\in\{1,\ldots,m\}$. By the definition of E, this implies that

$$xy\in E' \implies d(x,y)\leqslant\sum_{i=1}^{n} d(x_i,x_{i-1}) \leqslant m\frac{\varepsilon^2}{\sqrt{\log n}}\lesssim\varepsilon^2. \tag{15}$$

It is a standard fact (the equivalence between edge expansion and a Cheeger inequality) that for every $f : V \to L_1$ we have

$$\frac{1}{|E'|}\sum_{xy\in E'}\|f(x)-f(y)\|_1 \geqslant \frac{\alpha(V,E')}{2|V|^2}\sum_{x,y\in V}\|f(x)-f(y)\|_1. \tag{16}$$

For a proof of (16) see [59, Fact 2.1]: this is a simple consequence of the cut cone representation, i.e., Corollary 3.2, since the identity (3) shows that it suffices to prove (16) when $f(x) = \mathbf{1}_S(x)$ for some $S \subseteq V$, in which case the desired inequality follows immediately from the definition of the edge expansion $\alpha(V, E')$.

Since L_2 is isometric to a subset of L_1 (see, e.g., [71]), it follows from (16) and the fact that $\alpha(V, E') \gtrsim 1$ that

$$\begin{aligned}\varepsilon \overset{(15)}{\gtrsim} \frac{1}{|E'|}\sum_{xy\in E'}\sqrt{d(x,y)} &= \frac{1}{|E'|}\sum_{xy\in E'}\|v_x-v_y\|_2 \\ &\gtrsim \frac{1}{|V|^2}\sum_{x,y\in V}\|v_x-v_y\|_2 \gtrsim \frac{1}{n^2}\sum_{x,y\in V}\sqrt{d(x,y)}.\end{aligned} \tag{17}$$

Now comes the point where we use the assumption $\frac{1}{n^2}\sum_{x,y\in B(x_0,4)} d(x,y) \geqslant \frac{1}{4}$. Since for any $x, y \in B(x_0, 4)$ we have $d(x, y) \leqslant 8$, it follows that the number of pairs $(x, y) \in B(x_0, 4) \times B(x_0, 4)$ with $d(x, y) \geqslant 1/8$ is at least $n^2/64$. Since $|V| \geqslant (1-\varepsilon)|B(x_0, 4)|$, the number of such pairs which are also in $V \times V$ is at least $\frac{n^2}{64} - 3\varepsilon n^2 \gtrsim n^2$, provided ε is small enough. Thus $\frac{1}{n^2}\sum_{x,y\in V}\sqrt{d(x,y)} \gtrsim 1$, and (17) becomes a contradiction for small enough ε. □

Remark 4.8. The above proof of Theorem 4.7 used very little of the fact that d is a metric of negative type. In fact, all that was required was that d admits a quasisymmetric embedding into L_2; see [59].

It remains to say a few words about the proof of Theorem 4.4. Unfortunately, the present proof of this theorem is long and involved, and it relies on a variety of results from metric embedding theory. It would be of interest to obtain a simpler proof. Lemma 4.5 implies that Theorem 4.4 is a consequence of the following embedding result:

Theorem 4.9 ([5]). *Every n-point metric space of negative type embeds into Hilbert space with distortion $\lesssim (\log n)^{\frac{1}{2}+o(1)}$.*

Theorem 4.9 improves over the previously known [23] bound of $\lesssim (\log n)^{3/4}$ on the Euclidean distortion of n-point metric spaces of negative type. As we shall explain below, Theorem 4.9 is tight up to the $o(1)$ term.

The proof of Theorem 4.9 uses the following notion from [5]:

Definition 4.10 (Random zero-sets [5]). Fix $\Delta,\ \zeta > 0$, and $p \in (0,1)$. A metric space $(\mathscr{M}, d)$ is said to admit a random zero set at scale Δ, which is ζ-spreading with probability p, if there is a probability distribution μ over subsets $Z \subseteq \mathscr{M}$ such that $\mu(\{Z :\ y \in Z\ \wedge\ d(x,Z) \geqslant \Delta/\zeta\}) \geqslant p$ for every $x, y \in \mathscr{M}$ with $d(x,y) \geqslant \Delta$. We denote by $\zeta(\mathscr{M}; p)$ the least $\zeta > 0$ such that for every $\Delta > 0$, $\mathscr{M}$ admits a random zero set at scale Δ which is ζ-spreading with probability p.

The connection to metrics of negative type is due to the following theorem, which can be viewed as the main structural consequence of [6]. Its proof uses [6] in conjunction with two additional ingredients: an analysis of the algorithm of [6] due to [50], and a clever iterative application of the algorithm of [6], due to [23], while carefully reweighting points at each step.

Theorem 4.11 (Random zero sets for negative type metrics). *There exists a universal constant $p > 0$ such that any n-point metric space $(\mathscr{M}, d)$ of negative type satisfies $\zeta(\mathscr{M}; p) \lesssim \sqrt{\log n}$.*

Random zero sets are related to embeddings as follows. Fix $\Delta > 0$. Let $(\mathscr{M}, d)$ be a finite metric space, and fix $S \subseteq \mathscr{M}$. By the definition of $\zeta(S; p)$, there exists a distribution μ over subsets $Z \subseteq S$ such that for every $x, y \in S$ with $d(x,y) \geqslant \Delta$ we have $\mu(\{Z \subseteq S :\ y \in Z\ \wedge\ d(x,Z) \geqslant \Delta/\zeta(S;p)\}) \geqslant p$. Define $\varphi_{S,\Delta} : \mathscr{M} \to L_2(\mu)$ by $\varphi_{S,\Delta}(x) = d(x,Z)$. Then $\varphi_{S,\Delta}$ is 1-Lipschitz, and for every $x, y \in S$ with $d(x,y) \geqslant \Delta$,

$$\begin{aligned} \|\varphi_{S,\Delta}(x) - \varphi_{S,\Delta}(y)\|_{L_2(\mu)} &= \left(\int_{2^S} [d(x,Z) - d(y,Z)]^2 \, d\mu(Z)\right)^{1/2} \\ &\geqslant \frac{\Delta\sqrt{p}}{\zeta(S;p)}. \end{aligned} \tag{18}$$

The remaining task is to "glue" the mappings $\{\varphi_{S,\Delta} :\ \Delta > 0,\ S \subseteq \mathscr{M}\}$ to form an embedding of $\mathscr{M}$ into Hilbert space with the distortion claimed in Theorem 4.9. A key ingredient of the proof of Theorem 4.9 is the embedding method called "Measured Descent", that was developed in [48]. The results of [48] were stated as embedding theorems rather than a gluing procedure; the realization that a part of the arguments of [48] can be formulated explicitly as a general "gluing lemma" is due to [50]. In [5] it was necessary to enhance the Measured Descent technique in order to prove the following key theorem, which together with (18) and Theorem 4.11 implies Theorem 4.9. See also [4] for a different enhancement of Measured Descent, which also implies Theorem 4.9. The proof of Theorem 4.12 is quite intricate; we refer to [5] for the details.

Theorem 4.12. *Let $(\mathscr{M}, d)$ be an n-point metric space. Suppose that there is $\varepsilon \in [1/2, 1]$ such that for every $\Delta > 0$, and every subset $S \subseteq \mathscr{M}$, there exists*

a 1-Lipschitz map $\varphi_{S,\Delta} : \mathscr{M} \to L_2$ *with* $||\varphi_{S,\Delta}(x) - \varphi_{S,\Delta}(y)||_2 \gtrsim \Delta/(\log |S|)^{\varepsilon}$ *whenever* $x, y \in S$ *and* $d(x,y) \geqslant \Delta$. *Then* $c_2(\mathscr{M}) \lesssim (\log n)^{\varepsilon} \log\log n$.

The following corollary is an obvious consequence of Theorem 4.9, due to the fact that L_1 is a metric space of negative type.

Corollary 4.13. *Every* $X \subseteq L_1$ *embeds into* L_2 *with distortion* $\lesssim (\log |X|)^{\frac{1}{2}+o(1)}$.

We stated Corollary 4.13 since it is of special importance: in 1969, Enflo [34] proved that the Hamming cube, i.e., $\{0,1\}^k$ equipped with the metric induced from ℓ_1^k, has Euclidean distortion $\sqrt{k}$. Corollary 4.13 says that up to lower order factors, the Hamming cube is among the most non-Euclidean subset of L_1. There are very few known results of this type, i.e., (almost) sharp evaluations of the largest Euclidean distortion of an n-point subset of a natural metric space. A notable such result is Matoušek's theorem [57] that any n-point subset of the infinite binary tree has Euclidean distortion $\lesssim \sqrt{\log\log n}$, and consequently, due to [20], the same holds true for n-point subsets of, say, the hyperbolic plane. This is tight due to Bourgain's matching lower bound [16] for the Euclidean distortion of finite depth complete binary trees.

4.3.2. The Goemans-Linial conjecture. Theorem 4.4 is the best known approximation algorithm for the Sparsest Cut Problem (and Theorem 4.3 is the best known algorithm in the case of uniform demands). But, a comparison of Lemma 4.5 and Theorem 4.9 reveals a possible avenue for further improvement: Theorem 4.9 produces an embedding of negative type metrics into L_2 (for which the bound of Theorem 4.9 is sharp up to lower order factors), while for Lemma 4.5 all we need is an embedding into the larger space L_1. It was conjectured by Goemans and Linial (see [37, 55, 54] and [58, pg. 379–380]) that any finite metric space of negative type embeds into L_1 with distortion $\lesssim 1$. If true, this would yield, via the Goemans-Linial semidefinite relaxation, a constant factor approximation algorithm for Sparsest Cut.

As we shall see below, it turns out that the Goemans-Linial conjecture is false, and in fact there exist [30] arbitrarily large n-point metric spaces $\mathscr{M}_n$ of negative type for which $c_1(\mathscr{M}_n) \geqslant (\log n)^c$, where c is a universal constant. Due to the duality argument in Lemma 4.5, this means that the algorithm of Section 4.3 is doomed to make an error of at least $(\log n)^c$, i.e., there exist capacity and demand functions $C_n, D_n : \{1,\ldots,n\} \times \{1,\ldots,n\} \to [0,\infty)$ for which we have $M^{**}(C_n, D_n) \lesssim \Phi^*(C_n, D_n)/(\log n)^c$. Such a statement is referred to in the literature as the fact that the *integrality gap* of the Goemans-Linial semidefinite relaxation of Sparsest Cut is at least $(\log n)^c$.

4.3.3. Unique Games hardness and the Khot-Vishnoi integrality gap. Khot's Unique Games Conjecture [44] is that for every $\varepsilon > 0$ there exists a prime $p = p(\varepsilon)$ such that there is no polynomial time algorithm that, given $n \in \mathbb{N}$ and a system of m-linear equations in n-variables of the form $x_i - x_j = c_{ij}$

mod p for some $c_{ij} \in \mathbb{N}$, determines whether there exists an assignment of an integer value to each variable x_i such that at least $(1-\varepsilon)m$ of the equations are satisfied, or whether no assignment of such values can satisfy more than εm of the equations (if neither of these possibilities occur, then an arbitrary output is allowed). This formulation of the conjecture is due to [46], where it is shown that it is equivalent to the original formulation in [44]. The Unique Games Conjecture is by now a common assumption that has numerous applications in computational complexity; see the survey [45] (in this collection) for more information.

In [47, 24] it was shown that the existence of a polynomial time constant factor approximation algorithm for Sparsest Cut would refute the Unique Games Conjecture, i.e., one can use a polynomial time constant factor approximation algorithm for Sparsest Cut to solve in polynomial time the above algorithmic task for linear equations.

For a period of time in 2004, this computational hardness result led to a strange situation: either the complexity theoretic Unique Games Conjecture is true, or the purely geometric Goemans-Linial conjecture is true, but not both. In a remarkable tour de force, Khot and Vishnoi [47] delved into the proof of their hardness result and managed to construct from it a concrete family of arbitrarily large n-point metric spaces $\mathscr{M}_n$ of negative type for which $c_1(\mathscr{M}_n) \gtrsim (\log\log n)^c$, where c is a universal constant, thus refuting the Goemans-Linial conjecture. Subsequently, these Khot-Vishnoi metric spaces $\mathscr{M}_n$ were analyzed in [49], resulting in the lower bound $c_1(\mathscr{M}_n) \gtrsim \log\log n$. Further work in [32] yielded a $\gtrsim \log\log n$ integrality gap for Sparsest Cut with uniform demands, i.e., "average distortion" L_1 embeddings (in the sense of Lemma 4.6) of negative type metrics were ruled out as well.

4.3.4. The Bretagnolle, Dacunha-Castelle, Krivine theorem and invariant metrics on Abelian groups. A combination of Schoenberg's classical characterization [65] of metric spaces that are isometric to subsets of Hilbert space, and a theorem of Bretagnolle, Dacunha-Castelle and Krivine [18] (see also [70]), implies that if $p \in [1,2]$ and $(X, \|\cdot\|_X)$ is a separable Banach space such that the metric space $(X, \|x-y\|_X^{p/2})$ is isometric to a subset of Hilbert space, then X is (linearly) isometric to a subspace of L_p. Specializing to $p = 1$ we see that the Goemans-Linial conjecture is true for Banach spaces. With this motivation for the Goemans-Linial conjecture in mind, one notices that the Goemans-Linial conjecture is part of a natural one parameter family of conjectures which attempt to extend the theorem Bretagnolle, Dacunha-Castelle and Krivine to general metric spaces rather than Banach spaces: is it true that for $p \in [1,2)$ any metric space $(\mathscr{M}, d)$ for which $(\mathscr{M}, d^{p/2})$ is isometric to a subset of L_2 admits a bi-Lipschitz embedding into L_p? This generalized Goemans-Linial conjecture turns out to be false for all $p \in [1,2)$; our example based on the Heisenberg group furnishes counter-examples for all p.

It is also known that certain invariant metrics on Abelian groups satisfy the Goemans-Linial conjecture:

Theorem 4.14 ([10]). *Let G be a finite Abelian group, equipped with an invariant metric ρ. Suppose that $2 \leqslant m \in \mathbb{N}$ satisfies $mx = 0$ for all $x \in G$. Denote $D = c_2\left(G, \sqrt{\rho}\right)$. Then $c_1(G,\rho) \lesssim D^4 \log m$.*

It is an interesting open question whether the dependence on the exponent m of the group G in Theorem 4.14 is necessary. Can one construct a counter-example to the Goemans-Linial conjecture which is an invariant metric on the cyclic group C_n of order n? Or, is there for every $D \geqslant 1$ a constant $K(D)$ such that for every invariant metric ρ on C_n for which $c_2\left(G, \sqrt{\rho}\right) \leqslant D$ we have $c_1(G,\rho) \leqslant K(D)$?

One can view the above discussion as motivation for why one might consider the Heisenberg group as a potential counter-example to the Goemans-Linial conjecture. Assuming that we are interested in invariant metrics on groups, we wish to depart from the setting of Abelian groups or Banach spaces, and if at the same time we would like our example to have some useful analytic properties (such as invariance under rescaling and the availability of a group norm), the Heisenberg group suggests itself as a natural candidate. This plan is carried out in Section 5.

5. Embeddings of the Heisenberg Group

The purpose of this section is to discuss Theorem 1.1 and Theorem 1.2 from the introduction. Before doing so, we have an important item of unfinished business: relating the Heisenberg group to the Sparsest Cut Problem. We will do this in Section 5.1, following [52].

In preparation, we need to recall the Carnot-Carathéodory geometry of the continuous Heisenberg group $\mathbb{H}$, i.e., $\mathbb{R}^3$ equipped with the non-commutative product $(a,b,c)\cdot(a',b',c') = (a+a', b+b', c+c'+ab'-ba')$. Due to lack of space, this will have to be a crash course, and we refer to the relevant introductory sections of [29] for a more thorough discussion.

The identity element of $\mathbb{H}$ is $e = (0,0,0)$, and the inverse element of $(a,b,c) \in \mathbb{H}$ is $(-a,-b,-c)$. The center of $\mathbb{H}$ is the z-axis $\{0\} \times \{0\} \times \mathbb{R}$. For $g \in \mathbb{H}$ the *horizontal plane* at g is defined as $\mathbb{H}_g = g(\mathbb{R} \times \mathbb{R} \times \{0\})$. An affine line $L \subseteq \mathbb{H}$ is called a *horizontal line* if for some $g \in \mathbb{H}$ it passes through g and is contained in the affine plane $\mathbb{H}_g$. The standard scalar product $\langle \cdot, \cdot \rangle$ on $\mathbb{H}_e$ naturally induces a scalar product $\langle \cdot, \cdot \rangle_g$ on $\mathbb{H}_g$ by $\langle gx, gy \rangle_g = \langle x, y \rangle$. Consequently, we can define the Carnot-Carathéodory metric $d^{\mathbb{H}}$ on $\mathbb{H}$ by letting $d^{\mathbb{H}}(g,h)$ be the infimum of lengths of smooth curves $\gamma : [0,1] \to \mathbb{H}$ such that $\gamma(0) = g$, $\gamma(1) = h$ and for all $t \in [0,1]$ we have $\gamma'(t) \in H_{\gamma(t)}$ (and, the length of $\gamma'(t)$ is computed with respect to the scalar product $\langle \cdot, \cdot \rangle_{\gamma(t)}$). The *ball-box principle* (see [39]) implies that $d^{\mathbb{H}}\big((a,b,c),(a',b',c')\big)$ is bounded above and below by a constant

multiple of $|a-a'|+|b-b'|+\sqrt{|c-c'+ab'-ba'|}$. Moreover, since the integer grid $\mathbb{H}(\mathbb{Z})$ is a discrete cocompact subgroup of $\mathbb{H}$, the word metric d_W on $\mathbb{H}(\mathbb{Z})$ is bi-Lipschitz equivalent to the restriction of $d^{\mathbb{H}}$ to $\mathbb{H}(\mathbb{Z})$ (see, e.g, [19]). For $\theta > 0$ define the dilation operator $\delta_\theta : \mathbb{H} \to \mathbb{H}$ by $\delta_\theta(a,b,c) = (\theta a, \theta b, \theta^2 c)$. Then for all $g, h \in \mathbb{H}$ we have $d^{\mathbb{H}}(\delta_\theta(g), \delta_\theta(h)) = \theta d^{\mathbb{H}}(g,h)$. The Lebesgue measure $\mathscr{L}_3$ on $\mathbb{R}^3$ is a Haar measure of $\mathbb{H}$, and the volume of a $d^{\mathbb{H}}$-ball of radius r is proportional to r^4.

5.1. Heisenberg metrics with isometric L_p snowflakes. For every $(a,b,c) \in \mathbb{H}$ and $p \in [1,2)$, define

$$M_p(a,b,c) = \sqrt[4]{(a^2+b^2)^2+4c^2} \cdot \left(\cos\left(\frac{p}{2} \arccos\left(\frac{a^2+b^2}{\sqrt{(a^2+b^2)^2+4c^2}} \right) \right) \right)^{1/p}.$$

It was shown in [52] that M_p is a *group norm* on $\mathbb{H}$, i.e., for all $g, h \in \mathbb{H}$ and $\theta \geqslant 0$ we have $M_p(gh) \leqslant M_p(g) + M_p(h)$, $M_p(g^{-1}) = M_p(g)$ and $M_p(\delta_\theta(g)) = \theta M_p(g)$. Thus $d_p(g,h) \stackrel{\text{def}}{=} M_p(g^{-1}h)$ is a left-invariant metric on $\mathbb{H}$. The metric d_p is bi-Lipschitz equivalent to $d^{\mathbb{H}}$ with distortion of order $1/\sqrt{2-p}$ (see [52]). Moreover, it was shown in [52] that $(\mathbb{H}, d_p^{p/2})$ admits an isometric embedding into L_2. Thus, in particular, the metric space $(\mathbb{H}, d_1)$, which bi-Lipschitz equivalent to $(\mathbb{H}, d^{\mathbb{H}})$, is of negative type.

The fact that $(\mathbb{H}, d^{\mathbb{H}})$ does not admit a bi-Lipschitz embedding into L_p for any $1 \leqslant p < \infty$ will show that the generalized Goemans-Linial conjecture (see Section 4.3.4) is false. In particular, $(\mathbb{H}, d_1)$, and hence by a standard rescaling argument also $(\mathbb{H}(\mathbb{Z}), d_1)$, is a counter-example to the Goemans-Linial conjecture. Note that it is crucial here that we are dealing with the function space L_p rather than the sequence space ℓ_p, in order to use a compactness argument to deduce from this statement that there exist arbitrarily large n-point metric spaces $(\mathscr{M}_n, d)$ such that $(\mathscr{M}_n, d^{p/2})$ is isometric to a subset of L_2, yet $\lim_{n\to\infty} c_p(\mathscr{M}_n) = \infty$. The fact that this statement follows from non-embeddability into L_p is a consequence of a well known ultrapower argument (see [42]), yet for ℓ_p this statement is false (e.g., ℓ_2 does not admit a bi-Lipschitz embedding into ℓ_p, but all finite subsets of ℓ_2 embed isometrically into ℓ_p). Unfortunately, this issue creates substantial difficulties in the case of primary interest $p = 1$. In the reflexive range $p > 1$, or for a separable dual space such as ℓ_1 $(= c_0^*)$, the non-embeddability of $\mathbb{H}$ follows from a natural extension of a classical result of Pansu [61], as we explain in Section 5.2. This approach fails badly when it comes to embeddings into L_1: for this purpose a novel method of Cheeger and Kleiner [25] is needed, as described in Section 5.3.

5.2. Pansu differentiability. Let X be a Banach space and $f : \mathbb{H} \to X$. Following [61], f is said to have a Pansu derivative at $x \in \mathbb{H}$ if for every $y \in \mathbb{H}$ the limit $D_f^x(y) \stackrel{\text{def}}{=} \lim_{\theta\to 0} \big(f(x\delta_\theta(y)) - f(x)\big)/\theta$ exists, and $D_f^x : \mathbb{H} \to X$ is a group homomorphism, i.e., for all $y_1, y_2 \in \mathbb{H}$ we have $D_f^x(y_1 y_2^{-1}) = D_f^x(y_1) -$

$D_f^y(y_2)$. Pansu proved [61] that every $f : \mathbb{H} \to \mathbb{R}^n$ which is Lipschitz in the metric $d^{\mathbb{H}}$ is Pansu differentiable almost everywhere. It was observed in [52, 27] that this result holds true if the target space $\mathbb{R}^n$ is replaced by any Banach space with the Radon-Nikodým property, in particular X can be any reflexive Banach space such as L_p for $p \in (1, \infty)$, or a separable dual Banach space such as ℓ_1. As noted by Semmes [66], this implies that $\mathbb{H}$ does not admit a bi-Lipschitz embedding into any Banach space X with the Radon-Nikodým property: a bi-Lipschitz condition for f implies that at a point $x \in \mathbb{H}$ of Pansu differentiability, D_f^x is also bi-Lipschitz, and in particular a group isomorphism. But that's impossible since $\mathbb{H}$ is non-commutative, unlike the additive group of X.

5.3. Cheeger-Kleiner differentiability. Differentiability theorems fail badly when the target space is L_1, even for functions defined on $\mathbb{R}$; consider Aronszajn's example [3] of the "moving indicator function" $t \mapsto \mathbf{1}_{[0,t]} \in L_1$. For L_1-valued Lipschitz functions on $\mathbb{H}$, Cheeger and Kleiner [25, 28] developed an alternative differentiation theory, which is sufficiently strong to show that $\mathbb{H}$ does not admit a bi-Lipschitz embedding into L_1. Roughly speaking, a differentiation theorem states that in the infinitesimal limit, a Lipschitz mapping converges to a mapping that belongs to a certain "structured" subclass of mappings (e.g., linear mappings or group homomorphisms). The Cheeger-Kleiner theory shows that, in a sense that will be made precise below, L_1-valued Lipschitz functions on $\mathbb{H}$ are in the infinitesimal limit similar to Aronszajn's moving indicator.

For an open subset $U \subseteq \mathbb{H}$ let $\mathrm{Cut}(U)$ denote the space of (equivalences classes up to measure zero) of measurable subsets of U. Let $f : U \to L_1$ be a Lipschitz function. An infinitary variant of the cut-cone decomposition of Corollary 3.2 (see [25]) asserts that there exists a measure Σ_f on $\mathrm{Cut}(U)$, such that for all $x, y \in U$ we have $\|f(x) - f(y)\|_1 = \int_{\mathrm{Cut}(U)} |\mathbf{1}_E(x) - \mathbf{1}_E(y)| d\Sigma_f(E)$. The measure Σ_f is called the *cut measure* of f. The idea of Cheeger and Kleiner is to detect the "infinitesimal regularity" of f in terms of the infinitesimal behavior of the measure Σ_f; more precisely, in terms of the shape of the sets E in the support of Σ_f, after passing to an infinitesimal limit.

Theorem 5.1 (Cheeger-Kleiner differentiability theorem [25, 28]). *For almost every $x \in U$ there exists a measure Σ_f^x on $\mathrm{Cut}(\mathbb{H})$ such that for all $y, z \in \mathbb{H}$ we have*

$$\lim_{\theta \to 0} \frac{\|f(x\delta_\theta(y)) - f(x\delta_\theta(z))\|_1}{\theta} = \int_{\mathrm{Cut}(\mathbb{H})} |\mathbf{1}_E(y) - \mathbf{1}_E(z)| d\Sigma_f^x(E). \tag{19}$$

Moreover, the measure Σ_f^x is supported on affine half-spaces whose boundary is a vertical plane, i.e., a plane which isn't of the form $\mathbb{H}_g$ for some $g \in \mathbb{H}$ (equivalently, an inverse image, with respect to the orthogonal projection from $\mathbb{R}^3$ onto $\mathbb{R} \times \mathbb{R} \times \{0\}$, of a line in $\mathbb{R} \times \mathbb{R} \times \{0\}$).

Theorem 5.1 is incompatible with f being bi-Lipschitz, since the right hand side of (19) vanishes when y, z lie on the same coset of the center of $\mathbb{H}$, while if f is bi-Lipschitz the left hand side of (19) is at least a constant multiple of $d^{\mathbb{H}}(y, z)$.

5.4. Compression bounds for L_1 embeddings of the Heisenberg group. Theorem 1.1 and Theorem 1.2 are both a consequence of the following result from [29]:

Theorem 5.2 (Quantitative central collapse [29]). *There exists a universal constant $c \in (0,1)$ such that for every $p \in \mathbb{H}$, every 1-Lipschitz $f : B(p,1) \to L_1$, and every $\varepsilon \in \left(0, \frac{1}{4}\right)$, there exists $r \geqslant \varepsilon$ such that with respect to Haar measure, for at least half of the points $x \in B(p, 1/2)$, at least half of the points $(x_1, x_2) \in B(x,r) \times B(x,r)$ which lie on the same coset of the center satisfy:*

$$\|f(x_1) - f(x_2)\|_1 \leqslant \frac{d^{\mathbb{H}}(x_1, x_2)}{(\log(1/\varepsilon))^c}\,.$$

It isn't difficult to see that Theorem 5.2 implies Theorem 1.1 and Theorem 1.2. For example, in the setting of Theorem 1.1 we are given a bi-Lipschitz embedding $f : \{1, \ldots, n\}^3 \to L_1$, and using either the general extension theorem of [51] or a partition of unity argument, we can extend f to a Lipschitz (with respect to $d^{\mathbb{H}}$) mapping $\bar{f} : [1, n]^3 \to L_1$, whose Lipschitz constant is at most a constant multiple of the Lipschitz constant of f. Theorem 5.2 (after rescaling by n) produces a pair of points $y, z \in [1, n]^3$ of distance $\gtrsim \sqrt{n}$, whose distance is contracted under $\bar{f}$ by $\gtrsim (\log n)^c$. By rounding y, z to their nearest integer points in $\{1, \ldots, n\}^3$, we conclude that f itself must have bi-Lipschitz distortion $\gtrsim (\log n)^c$. The deduction of Theorem 1.2 from Theorem 5.2 is just as simple; see [29].

Theorem 5.2 is a quantitative version of Theorem 5.1, in the sense it gives a definite lower bound on the macroscopic scale at which a given amount of collapse of cosets of the center, as exhibited by the differentiation result (19), occurs. As explained in [29, Rem. 2.1], one cannot hope in general to obtain rate bounds in differentiation results such as (19). Nevertheless, there are situations where "quantitative differentiation results" have been successfully proved; important precursors of Theorem 5.2 include the work of Bourgain [17], Jones [43], Matoušek [57], and Bates, Johnson, Lindenstrauss, Preiss, Schechtman [13]. Specifically, we should mention that Bourgain [17] obtained a lower bound on $\varepsilon > 0$ such that any embedding of an ε-net in a unit ball of an n-dimensional normed space X into a normed space Y has roughly the same distortion as the distortion required to embed all of X into Y, and Matoušek [57], in his study of embeddings of trees into uniformly convex spaces, obtained quantitative bounds on the scale at which "metric differentiation" is almost achieved, i.e., a scale at which discrete geodesics are mapped by a Lipschitz function to "almost geodesics". These earlier results are in the spirit of Theorem 5.2, though the proof of Theorem 5.2 in [29] is substantially more involved.

We shall now say a few words on the proof of Theorem 5.2; for lack of space this will have to be a rough sketch, so we refer to [29] for more details, as well as to the somewhat different presentation in [30]. Cheeger and Kleiner obtained two different proofs of Theorem 5.1. The first proof [25] started with the important observation that the fact that f is Lipschitz forces the cut measure Σ_f to be supported on sets with additional regularity, namely sets of *finite perimeter*. Moreover, there is a definite bound on the *total perimeter*: $\int_{\mathrm{Cut}(U)} \mathrm{PER}(E, B(p,1)) d\Sigma_f(E) \lesssim 1$, where $\mathrm{PER}(E, B(p,1))$ denotes the perimeter of E in the ball $B(p,1)$ (we refer to the book [2], and the detailed explanation in [25, 29] for more information on these notions). Theorem 5.2 is then proved in [25] via an appeal to results [35, 36] on the infinitesimal structure of sets of finite perimeter in $\mathbb{H}$. A different proof of Theorem 5.2 was found in [28]. It is based on the notion of *metric differentiation*, which is used in [28] to reduce the problem to mappings $f : \mathbb{H} \to L_1$ for which the cut measure is supported on *monotone sets*, i.e., sets $E \subseteq \mathbb{H}$ such that for every horizontal line L, up to a set of measure zero, both $L \cap E$ and $L \cap (\mathbb{H} \setminus E)$ are either empty or subrays of L. A non-trivial classification of monotone sets is then proved in [28]: such sets are up to measure zero half-spaces.

This second proof of Theorem 5.2 avoids completely the use of perimeter bounds. Nevertheless, the starting point of the proof of Theorem 5.2 can be viewed as a hybrid argument, which incorporates both perimeter bounds, and a new classification of *almost* monotone sets. The quantitative setting of Theorem 5.2 leads to issues that do not have analogues in the non-quantitative proofs (e.g., the approximate classification results of "almost" monotone sets in balls cannot be simply that such sets are close to half-spaces in the entire ball; see [29, Example 9.1]).

In order to proceed we need to quantify the extent to which a set $E \subseteq B(x,r)$ is monotone. For a horizontal line $L \subseteq \mathbb{H}$ define the non-convexity $\mathrm{NC}_{B(x,r)}(E, L)$ of (E, L) on $B(x,r)$ as the infimum of $\int_{L \cap B(x,r)} |\mathbf{1}_I - \mathbf{1}_{E \cap L \cap B_r(x)}| d\mathcal{H}^1_L$ over all sub-intervals $I \subseteq L \cap B_r(x)$. Here $\mathcal{H}^1_L$ is the 1-dimensional Hausdorff measure on L (induced from the metric $d^{\mathbb{H}}$). The non-monotonicity of (E, L) on $B(x,r)$ is defined to be $\mathrm{NM}_{B(x,r)}(E, L) \stackrel{\mathrm{def}}{=} \mathrm{NC}_{B(x,r)}(E, L) + \mathrm{NC}_{B(x,r)}(\mathbb{H} \setminus E, L)$. The total non-monotonicity of E on $B(x,r)$ is defined as:

$$\mathrm{NM}_{B(x,r)}(E) \stackrel{\mathrm{def}}{=} \frac{1}{r^4} \int_{\mathrm{lines}(B(x,r))} \mathrm{NM}_{B(x,r)}(E, L) d\mathcal{N}(L),$$

where $\mathrm{lines}(U)$ denotes the set of horizontal lines in $\mathbb{H}$ which intersect U, and $\mathcal{N}$ is the left invariant measure on $\mathrm{lines}(\mathbb{H})$, normalized so that the measure of $\mathrm{lines}(B(e,1))$ is 1.

The following stability result for monotone sets constitutes the bulk of [29]:

Theorem 5.3. *There exists a universal constant $a > 0$ such that if a measurable set $E \subseteq B(x,r)$ satisfies $\mathrm{NM}_{B(x,r)}(E) \leqslant \varepsilon^a$ then there exists a half-space*

$\mathcal{P}$ *such that*

$$\frac{\mathscr{L}_3\left((E \cap B_{\varepsilon r}(x)) \triangle \mathcal{P}\right)}{\mathscr{L}_3(B_{\varepsilon r}(x))} < \varepsilon^{1/3}.$$

Perimeter bounds are used in [29, 30] for two purposes. The first is finding a controlled scale r such that at most locations, apart from a certain collection of cuts, the mass of Σ_f is supported on subsets which satisfy the assumption of Theorem 5.3 (see [30, Sec. 9]). But, the excluded cuts may have infinite measure with respect to Σ_f. Nonetheless, using perimeter bounds once more, together with the isoperimetric inequality in $\mathbb{H}$ (see [60, 21]), it is shown that their contribution to the metric is negligibly small (see [30, Sec. 8]).

By Theorem 5.3, it remains to deal with the situation where all the cuts in the support of Σ_f are close to half-spaces: note that we are not claiming in Theorem 5.3 that the half-space is vertical. Nevertheless, a simple geometric argument shows that even in the case of cut measures that are supported on general (almost) half-spaces, the mapping f must significantly distort some distances. The key point here is that if the cut measure is actually supported on half spaces, then it follows (after the fact) that for *every affine* line L, if $x_1, x_2, x_3 \in L$ and x_2 lies between x_1 and x_3 then $\|f(x_1) - f(x_3)\|_1 = \|f(x_1) - f(x_2)\|_1 + \|f(x_2) - f(x_3)\|_1$. But if L is vertical then $d^{\mathbb{H}}|_L$ is bi-Lipschitz to the *square root* of the difference of the z-coordinates, and it is trivial to verify that this metric on L is not bi-Lipschitz equivalent to a metric on L satisfying this additivity condition. For the details of (a quantitative version of) this final step of the argument see [30, Sec. 10].

References

[1] A. Agrawal, P. Klein, R. Ravi, and S. Rao. Approximation through multicommodity flow. In *31st Annual Symposium on Foundations of Computer Science*, pages 726–737. IEEE Computer Soc., Los Alamitos, CA, 1990.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, 2000.

[3] N. Aronszajn. Differentiability of Lipschitzian mappings between Banach spaces. *Studia Math.*, 57(2):147–190, 1976.

[4] S. Arora, J. R. Lee, and A. Naor. Fréchet embeddings of negative type metrics. *Discrete Comput. Geom.*, 38(4):726–739, 2007.

[5] S. Arora, J. R. Lee, and A. Naor. Euclidean distortion and the sparsest cut. *J. Amer. Math. Soc.*, 21(1):1–21 (electronic), 2008.

[6] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 222–231 (electronic), New York, 2004. ACM.

[7] G. Arzhantseva, C. Drutu, and M. Sapir. Compression functions of uniform embeddings of groups into Hilbert and Banach spaces. *J. Reine Angew. Math.*, 633:213–235, 2009.

[8] P. Assouad. Plongements Lipschitziens dans $\mathbf{R}^n$. *Bull. Soc. Math. France*, 111(4):429–448, 1983.

[9] Y. Aumann and Y. Rabani. An $O(\log k)$ approximate min-cut max-flow theorem and approximation algorithm. *SIAM J. Comput.*, 27(1):291–301 (electronic), 1998.

[10] T. Austin, A. Naor, and A. Valette. The Euclidean distortion of the lamplighter group. Preprint available at http://arxiv.org/abs/0705.4662. To appear in Discrete Comput. Geom., 2007.

[11] D. Avis and M. Deza. The cut cone, L^1 embeddability, complexity, and multicommodity flows. *Networks*, 21(6):595–617, 1991.

[12] K. Ball. Isometric embedding in l_p-spaces. *European J. Combin.*, 11(4):305–311, 1990.

[13] S. Bates, W. B. Johnson, J. Lindenstrauss, D. Preiss, and G. Schechtman. Affine approximation of Lipschitz functions and nonlinear quotients. *Geom. Funct. Anal.*, 9(6):1092–1127, 1999.

[14] Y. Benyamini and J. Lindenstrauss. *Geometric nonlinear functional analysis. Vol. 1*, volume 48 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2000.

[15] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.*, 52(1-2):46–52, 1985.

[16] J. Bourgain. The metrical interpretation of superreflexivity in Banach spaces. *Israel J. Math.*, 56(2):222–230, 1986.

[17] J. Bourgain. Remarks on the extension of Lipschitz maps defined on discrete sets and uniform homeomorphisms. In *Geometrical aspects of functional analysis (1985/86)*, volume 1267 of *Lecture Notes in Math.*, pages 157–167. Springer, Berlin, 1987.

[18] J. Bretagnolle, D. Dacunha–Castelle, and J.-L. Krivine. Lois stables et espaces L^p. *Ann. Inst. H. Poincaré Sect. B (N.S.)*, 2:231–259, 1965/1966.

[19] D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[20] S. Buyalo and V. Schroeder. Embedding of hyperbolic spaces in the product of trees. *Geom. Dedicata*, 113:75–93, 2005.

[21] L. Capogna, D. Danielli, and N. Garofalo. The geometric Sobolev embedding for vector fields and the isoperimetric inequality. *Comm. Anal. Geom.*, 2(2):203–215, 1994.

[22] S. Chawla. Sparsest cut. In M.-Y. Kao, editor, *Encyclopedia of Algorithms*. Springer, 2008.

[23] S. Chawla, A. Gupta, and H. Räcke. Embeddings of negative-type metrics and an improved approximation to generalized sparsest cut. *ACM Trans. Algorithms*, 4(2):Art. 22, 18, 2008.

[24] S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, and D. Sivakumar. On the hardness of approximating multicut and sparsest-cut. *Comput. Complexity*, 15(2):94–114, 2006.

[25] J. Cheeger and B. Kleiner. Differentiating maps into L^1 and the geometry of BV functions. To appear in Ann. Math., preprint available at http://arxiv.org/abs/math/0611954, 2006.

[26] J. Cheeger and B. Kleiner. Generalized differentiation and bi-Lipschitz nonembedding in L^1. *C. R. Math. Acad. Sci. Paris*, 343(5):297–301, 2006.

[27] J. Cheeger and B. Kleiner. On the differentiability of Lipschitz maps from metric measure spaces to Banach spaces. In *Inspired by S. S. Chern*, volume 11 of *Nankai Tracts Math.*, pages 129–152. World Sci. Publ., Hackensack, NJ, 2006.

[28] J. Cheeger and B. Kleiner. Metric differentiation, monotonicity and maps to L^1. Preprint available at http://arxiv.org/abs/0907.3295, 2009.

[29] J. Cheeger, B. Kleiner, and A. Naor. Compression bounds for Lipschitz maps from the Heisenberg group to L_1. Preprint, 2009. http://arxiv.org/abs/0910.2026.

[30] J. Cheeger, B. Kleiner, and A. Naor. A $(\log n)^{\Omega(1)}$ integrality gap for the Sparsest Cut SDP. In *Proceedings of 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2009)*, pages 555–564, 2009.

[31] J. Chuzhoy and S. Khanna. Polynomial flow-cut gaps and hardness of directed cut problems [extended abstract]. In *STOC'07—Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pages 179–188. ACM, New York, 2007.

[32] N. R. Devanur, S. A. Khot, R. Saket, and N. K. Vishnoi. Integrality gaps for sparsest cut and minimum linear arrangement problems. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 537–546. ACM, New York, 2006.

[33] M. M. Deza and M. Laurent. *Geometry of cuts and metrics*, volume 15 of *Algorithms and Combinatorics*. Springer–Verlag, Berlin, 1997.

[34] P. Enflo. On the nonexistence of uniform homeomorphisms between L_p-spaces. *Ark. Mat.*, 8:103–105 (1969), 1969.

[35] B. Franchi, R. Serapioni, and F. Serra Cassano. Rectifiability and perimeter in the Heisenberg group. *Math. Ann.*, 321(3):479–531, 2001.

[36] B. Franchi, R. Serapioni, and F. Serra Cassano. On the structure of finite perimeter sets in step 2 Carnot groups. *J. Geom. Anal.*, 13(3):421–466, 2003.

[37] M. X. Goemans. Semidefinite programming in combinatorial optimization. *Math. Programming*, 79(1–3, Ser. B):143–161, 1997. Lectures on mathematical programming (ismp97) (Lausanne, 1997).

[38] M. Gromov. Asymptotic invariants of infinite groups. In *Geometric group theory, Vol. 2 (Sussex, 1991)*, volume 182 of *London Math. Soc. Lecture Note Ser.*, pages 1–295. Cambridge Univ. Press, Cambridge, 1993.

[39] M. Gromov. Carnot–Carathéodory spaces seen from within. In *Sub-riemannian geometry*, Progr. in Math., pages 79–323. Birkhäuser, Basel, 1996.

[40] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2 of *Algorithms and Combinatorics*. Springer–Verlag, Berlin, second edition, 1993.

[41] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, pages 534–543. IEEE Computer Society, 2003.

[42] S. Heinrich. Ultraproducts in Banach space theory. *J. Reine Angew. Math.*, 313:72–104, 1980.

[43] P. W. Jones. Lipschitz and bi-Lipschitz functions. *Rev. Mat. Iberoamericana*, 4(1):115–121, 1988.

[44] S. Khot. On the power of unique 2-prover 1-round games. In *Proceedings of the Thirty–Fourth Annual ACM Symposium on Theory of Computing*, pages 767–775 (electronic), New York, 2002. ACM.

[45] S. Khot. Inapproximability of NP-complete problems, discrete Fourier analysis, and geometry. To appear in *Proceedings of the International Congress of Mathematicians, (Hyderabad, 2010)*, 2010.

[46] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for MAX–CUT and other 2-variable CSPs? *SIAM J. Comput.*, 37(1):319–357 (electronic), 2007.

[47] S. Khot and N. Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into ℓ_1. In *Proceedings of the 46th Annual IEEE Conference on Foundations of Computer Science (FOCS 2005)*, pages 53–62, 2005.

[48] R. Krauthgamer, J. R. Lee, M. Mendel, and A. Naor. Measured descent: a new embedding method for finite metrics. *Geom. Funct. Anal.*, 15(4):839–858, 2005.

[49] R. Krauthgamer and Y. Rabani. Improved lower bounds for embeddings into L_1. In *Proceedings of the Seventeenth Annual ACM–SIAM Symposium on Discrete Algorithms*, pages 1010–1017, New York, 2006. ACM.

[50] J. R. Lee. On distance scales, embeddings, and efficient relaxations of the cut cone. In *Proceedings of the Sixteenth Annual ACM–SIAM Symposium on Discrete Algorithms*, pages 92–101 (electronic), New York, 2005. ACM.

[51] J. R. Lee and A. Naor. Extending Lipschitz functions via random metric partitions. *Invent. Math.*, 160(1):59–95, 2005.

[52] J. R. Lee and A. Naor. L_p metrics on the Heisenberg group and the Goemans–Linial conjecture. In *Proceedings of 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006)*, pages 99–108. IEEE Computer Society, 2006.

[53] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *J. ACM*, 46(6):787–832, 1999.

[54] N. Linial. Finite metric-spaces—combinatorics, geometry and algorithms. In *Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002)*, pages 573–586, Beijing, 2002. Higher Ed. Press.

[55] N. Linial. Squared ℓ_2 metrics into ℓ_1. In *Open problems on embeddings of finite metric spaces, edited by J. Matoušek*, page 5. 2002.

[56] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[57] J. Matoušek. On embedding trees into uniformly convex Banach spaces. *Israel J. Math.*, 114:221–237, 1999.

[58] J. Matoušek. *Lectures on discrete geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer–Verlag, New York, 2002.

[59] A. Naor, Y. Rabani, and A. Sinclair. Quasisymmetric embeddings, the observable diameter, and expansion properties of graphs. *J. Funct. Anal.*, 227(2):273–303, 2005.

[60] P. Pansu. Une inégalité isopérimétrique sur le groupe de Heisenberg. *C. R. Acad. Sci. Paris Sér. I Math.*, 295(2):127–130, 1982.

[61] P. Pansu. Métriques de Carnot–Carathéodory et quasiisométries des espaces symétriques de rang un. *Ann. of Math. (2)*, 129(1):1–60, 1989.

[62] S. D. Pauls. The large scale geometry of nilpotent Lie groups. *Comm. Anal. Geom.*, 9(5):951–982, 2001.

[63] Y. Rabinovich. On average distortion of embedding metrics into the line. *Discrete Comput. Geom.*, 39(4):720–733, 2008.

[64] S. Rao. Small distortion and volume preserving embeddings for planar and Euclidean metrics. In *Proceedings of the Fifteenth Annual Symposium on Computational Geometry (Miami Beach, FL, 1999)*, pages 300–306 (electronic), New York, 1999. ACM.

[65] I. J. Schoenberg. Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, 44(3):522–536, 1938.

[66] S. Semmes. On the nonexistence of bi-Lipschitz parameterizations and geometric problems about A_∞-weights. *Rev. Mat. Iberoamericana*, 12(2):337–410, 1996.

[67] F. Shahrokhi and D. W. Matula. The maximum concurrent flow problem. *J. Assoc. Comput. Mach.*, 37(2):318–334, 1990.

[68] D. B. Shmoys. Cut problems and their application to divide-and-conquer. In *Approximation Algorithms for NP-hard Problems, (D.S. Hochbaum, ed.)*, pages 192–235. PWS, 1997.

[69] R. Tessera. Quantitative property A, Poincaré inequalities, L^p-compression and L^p-distortion for metric measure spaces. *Geom. Dedicata*, 136:203–220, 2008.

[70] J. H. Wells and L. R. Williams. *Embeddings and extensions in analysis*. Springer–Verlag, New York, 1975. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 84.

[71] P. Wojtaszczyk. *Banach spaces for analysts*, volume 25 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1991.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Non-asymptotic Theory of Random Matrices: Extreme Singular Values

Mark Rudelson* and Roman Vershynin†

Abstract

The classical random matrix theory is mostly focused on asymptotic spectral properties of random matrices as their dimensions grow to infinity. At the same time many recent applications from convex geometry to functional analysis to information theory operate with random matrices in fixed dimensions. This survey addresses the non-asymptotic theory of extreme singular values of random matrices with independent entries. We focus on recently developed geometric methods for estimating the hard edge of random matrices (the smallest singular value).

Mathematics Subject Classification (2010). Primary 60B20; Secondary 46B09

Keywords. Random matrices, singular values, hard edge, Littlewood-Offord problem, small ball probability

1. Asymptotic and Non-asymptotic Problems on Random Matrices

Since its inception, random matrix theory has been mostly preoccupied with asymptotic properties of random matrices as their dimensions grow to infinity. A foundational example of this nature is *Wigner's semicircle law* [96]. It applies to a family of $n \times n$ symmetric matrices A_n whose entries on and above the diagonal are independent standard normal random variables. In the limit as the dimension n grows to infinity, the spectrum of the normalized matrices $\frac{1}{\sqrt{n}} A_n$ is distributed according to the semicircle law with density $\frac{1}{2\pi}\sqrt{4-x^2}$

*Partially supported by NSF grant DMS FRG 0652684.
Department of Mathematics, University of Missouri-Columbia, Columbia, Missouri, U.S.A. E-mail: rudelsonm@missouri.edu.
†Partially supported by NSF grant DMS FRG 0918623.
Department of Mathematics, University of Michigan, Ann Arbor, Michigan, U.S.A. E-mail: romanv@umich.edu.

supported on the interval $[-2, 2]$. Precisely, if we denote by $S_n(z)$ the number of eigenvalues of $\frac{1}{\sqrt{n}} A_n$ that are smaller than z, then for every $z \in \mathbb{R}$ one has

$$\frac{S_n(z)}{n} \to \frac{1}{2\pi} \int_{-\infty}^{z} (4 - x^2)_+^{1/2} \, dx \quad \text{almost surely as } n \to \infty.$$

In a similar way, *Marchenko-Pastur law* [55] governs the limiting spectrum of $n \times n$ Wishart matrices $W_{N,n} = A^* A$, where $A = A_{N,n}$ is an $N \times n$ *random Gaussian matrix* whose entries are independent standard normal random variables. As the dimensions N, n grow to infinity while the aspect ratio n/N converges to a non-random number $y \in (0, 1]$, the spectrum of the normalized Wishart matrices $\frac{1}{N} W_{N,n}$ is distributed according to the Marchenko-Pastur law with density $\frac{1}{2\pi xy} \sqrt{(b-x)(x-a)}$ supported on $[a, b]$ where $a = (1 - \sqrt{y})^2$, $b = (1 + \sqrt{y})^2$. The meaning of the convergence is similar to the one in Wigner's semicircle law.

It is widely believed that phenomena typically observed in asymptotic random matrix theory are *universal*, that is independent of the particular distribution of the entries of random matrices. By analogy with classical probability, when we work with independent standard normal random variables Z_i, we know that their normalized sum $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i$ is again a standard normal random variable. This simple but useful fact becomes significantly more useful when we learn that it is asymptotically universal. Indeed, The Central Limit Theorem states that if instead of normal distribution Z_i have general identical distribution with zero mean and unit variance, the normalized sum S_n will still converge (in distribution) to the standard normal random variable as $n \to \infty$. In random matrix theory, universality has been established for many results. In particular, Wigner's semicircle law and Marchenko-Pastur law are known to be universal – like the Central Limit Theorem, they hold for arbitrary distribution of entries with zero mean and unit variance (see [60, 6] for semi-circle law and [95, 5] for Marchenko-Pastur law).

Asymptotic random matrix theory offers remarkably precise predictions as dimension grows to infinity. At the same time, sharpness at infinity is often counterweighted by lack of understanding of what happens in finite dimensions. Let us briefly return to the analogy with the Central Limit Theorem. One often needs to estimate the sum of independent random variables S_n with fixed number of terms n rather than in the limit $n \to \infty$. In this situation one may turn to Berry-Esseen's theorem which quantifies deviations of the distribution of S_n from that of the standard normal random variable Z. In particular, if $\mathbb{E}|Z_1|^3 = M < \infty$ then

$$|\mathbb{P}(S_n \le z) - \mathbb{P}(Z \le z)| \le \frac{C}{1 + |z|^3} \cdot \frac{M}{\sqrt{n}}, \quad z \in \mathbb{R}, \tag{1.1}$$

where C is an absolute constant [11, 23]. Notwithstanding the optimality of Berry-Esseen inequality (1.1), one can still hope for something better than the

polynomial bound on the probability, especially in view of the super-exponential tail of the limiting normal distribution: $\mathbb{P}(|Z| > z) \lesssim \exp(-z^2/2)$. Better estimates would indeed emerge in the form of *exponential deviation inequalities* [61, 47], but this would only happen when we drop explicit comparisons to the limiting distribution and study the tails of S_n by themselves. In the simplest case, when Z_i are i.i.d. mean zero random variables bounded in absolute value by 1, one has

$$\mathbb{P}(|S_n| > z) \le 2\exp(-cz^2), \quad z \ge 0, \tag{1.2}$$

where c is a positive absolute constant. Such exponential deviation inequalities, which are extremely useful in a number of applications, are non-asymptotic results whose asymptotic prototype is the Central Limit Theorem.

A similar non-asymptotic viewpoint can be adopted in random matrix theory. One would then study spectral properties of random matrices of fixed dimensions. Non-asymptotic results on random matrices are in demand in a number of today's applications that operate in high but fixed dimensions. This usually happens in statistics where one analyzes data sets with a large but fixed number of parameters, in geometric functional analysis where one works with random operators on finite-dimensional spaces (whose dimensions are large but fixed), in signal processing where the signal is randomly sampled in many but fixed number of points, and in various other areas of science and engineering.

This survey is mainly focused on the non-asymptotic theory of the extreme singular values of random matrices (equivalently, the extreme eigenvalues of sample covariance matrices) where significant progress was made recently. In Section 2 we review estimates on the largest singular value (the soft edge). The more difficult problem of estimating the smallest singular value (the hard edge) is discussed in Section 3, and its connection with the Littlewood-Offord problem in additive combinatorics is the content of Section 4. In Section 5 we discuss several applications of non-asymptotic results to the circular law in asymptotic random matrix theory, to restricted isometries in compressed sensing, and to Kashin's subspaces in geometric functional analysis.

This paper is by no means a comprehensive survey of the area but rather a tutorial. Sketches of some arguments are included in order to give the reader a flavor of non-asymptotic methods. To do this more effectively, we state most theorems in simplified form (e.g. always over the field $\mathbb{R}$); the reader will find full statements in the original papers. Also, we had to completely omit several important directions. These include random symmetric matrices which were the subject of the recent survey by Ledoux [48] and random matrices with independent columns, see in particular [1, 94]. The reader is also encouraged to look at the comprehensive survey [19] on some geometric aspects of random matrix theory.

2. Extreme Singular Values

Geometric nature of extreme singular values The non-asymptotic viewpoint in random matrix theory is largely motivated by geometric problems in high dimensional Euclidean spaces. When we view an $N \times n$ matrix A as a linear operator $\mathbb{R}^n \to \mathbb{R}^N$, we may want first of all to control its magnitude by placing useful upper and lower bounds on A. Such bounds are conveniently provided by the smallest and largest singular values of A denoted $s_{\min}(A)$ and $s_{\max}(A)$; recall that the singular values are by definition the eigenvalues of $|A| = \sqrt{A^*A}$.

The geometric meaning of the extreme singular values can be clear by considering the best possible factors m and M in the two-sided inequality

$$m\|x\|_2 \le \|Ax\|_2 \le M\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

The largest m and the smallest M are precisely the extreme singular values $s_{\min}(A)$ and $s_{\max}(A)$ respectively. They control the distortion of the Euclidean geometry under the action of the linear transformation A; the distance between any two points in $\mathbb{R}^n$ can increase by at most the factor $s_{\max}(A)$ and decrease by at most the factor $s_{\max}(A)$. The extreme singular values are clearly related to the operator norms of the linear operators A and A^{-1} acting between Euclidean spaces: $s_{\max}(A) = \|A\|$ and if A is invertible then $s_{\min}(A) = 1/\|A^{-1}\|$.

Understanding the behavior of extreme singular values of random matrices is needed in many applications. In numerical linear algebra, the *condition number* $\kappa(A) = s_{\max}(A)/s_{\min}(A)$ often serves as a measure of stability of matrix algorithms. Geometric functional analysis employs probabilistic constructions of linear operators as random matrices, and the success of these constructions often depends on good bounds on the norms of these operators and their inverses. Applications of different nature arise in statistics from the analysis of *sample covariance matrices* A^*A, where the rows of A are formed by N independent samples of some unknown distribution in $\mathbb{R}^n$. Some other applications are discussed in Section 5.

Asymptotic behavior of extreme singular values We first turn to the asymptotic theory for the extreme singular values of random matrices with independent entries (and with zero mean and unit variance for normalization purposes). From Marchenko-Pastur law we know that most singular values of such random $N \times n$ matrix A lie in the interval $[\sqrt{N} - \sqrt{n}, \sqrt{N} + \sqrt{n}]$. Under mild additional assumptions, it is actually true that *all* singular values lie there, so that asymptotically we have

$$s_{\min}(A) \sim \sqrt{N} - \sqrt{n}, \quad s_{\max}(A) \sim \sqrt{N} + \sqrt{n}. \tag{2.1}$$

This fact is universal and it holds for general distributions. This was established for $s_{\max}(A)$ by Geman [29] and Yin, Bai and Krishnaiah [97]. For $s_{\min}(A)$, Silverstein [71] proved this for Gaussian random matrices, and Bai and Yin [8] gave a unified treatment of both extreme singular values for general distributions:

Theorem 2.1 (Convergence of extreme singular values, see [8]). *Let $A = A_{N,n}$ be an $N \times n$ random matrix whose entries are independent copies of some random variable with zero mean, unit variance, and finite fourth moment. Suppose that the dimensions N and n grow to infinity while the aspect ratio n/N converges to some number $y \in (0, 1]$. Then*

$$\frac{1}{\sqrt{N}}\, s_{\min}(A) \to 1 - \sqrt{y}, \quad \frac{1}{\sqrt{N}}\, s_{\max}(A) \to 1 + \sqrt{y} \quad \text{almost surely.}$$

Moreover, without the fourth moment assumption the sequence $\frac{1}{\sqrt{N}}\, s_{\max}(A)$ is almost surely unbounded [7].

The limiting distribution of the extreme singular values is known and universal. It is given by the *Tracy-Widom law* whose cumulative distribution function is

$$F_1(x) = \exp\Big(- \int_x^\infty \big[u(s) + (s-x)u^2(s)\big]\, ds\Big), \tag{2.2}$$

where $u(s)$ is the solution to the Painlevè II equation $u'' = 2u^3 + su$ with the asymptotic $u(s) \sim \frac{1}{2\sqrt{\pi} s^{1/4}} \exp(-\frac{2}{3} s^{3/2})$ as $s \to \infty$. The occurrence of Tracy-Widom law in random matrix theory and several other areas was the subject of an ICM 2002 talk of Tracy and Widom [91]. This law was initially discovered for the largest eigenvalue of a Gaussian symmetric matrix [89, 90]. For the largest singular values of random matrices with independent entries it was established by Johansson [37] and Johnstone [39] in the Gaussian case, and by Soshnihikov [74] for more general distributions. For the smallest singular value, the corresponding result was recently obtained in a recent work Feldheim and Sodin [25] who gave a unified treatment of both extreme singular values. These results are known under a somewhat stronger subgaussian moment assumption on the entries a_{ij} of A, which requires their distribution to decay as fast as the normal random variable:

Definition 2.2 (Subgaussian random variables). A random variable X is *subgaussian* if there exists $K > 0$ called the subgaussian moment of X such that

$$\mathbb{P}(|X| > t) \le 2e^{-t^2/K^2} \quad \text{for } t > 0.$$

Examples of subgaussian random variables include normal random variables, ± 1-valued, and generally, all bounded random variables. The subgaussian assumption is equivalent to the moment growth condition $(\mathbb{E}|X|^p)^{1/p} = O(\sqrt{p})$ as $p \to \infty$.

Theorem 2.3 (Limiting distribution of extreme singular values, see [25]). *Let $A = A_{N,n}$ be an $N \times n$ random matrix whose entries are independent and identically distributed subgaussian random variables with zero mean and unit variance. Suppose that the dimensions N and n grow to infinity while the aspect ratio n/N stays uniformly bounded by some number $y \in (0, 1)$. Then the*

normalized extreme singular values

$$\frac{s_{\min}(A)^2 - (\sqrt{N} - \sqrt{n})^2}{(\sqrt{N} - \sqrt{n})(1/\sqrt{n} - 1/\sqrt{N})^{1/3}} \quad \textit{and} \quad \frac{s_{\max}(A)^2 - (\sqrt{N} + \sqrt{n})^2}{(\sqrt{N} + \sqrt{n})(1/\sqrt{n} + 1/\sqrt{N})^{1/3}}$$

converge in distribution to the Tracy-Widom law (2.2).

Non-asymptotic behavior of extreme singular values It is not entirely clear to what extent the limiting behavior of the extreme singular values such as asymptotics (2.1) manifests itself in fixed dimensions. Given the geometric meaning of the extreme singular values, our interest generally lies in establishing correct upper bounds on $s_{\max}(A)$ and lower bounds on $s_{\min}(A)$. We start with a folklore observation which yields the correct bound $s_{\max}(A) \lesssim \sqrt{N} + \sqrt{n}$ up to an absolute constant factor. The proof is a basic instance of an *ε-net argument*, a technique proved to be very useful in geometric functional analysis.

Proposition 2.4 (Largest singular value of subgaussian matrices: rough bound). *Let A be an $N \times n$ random matrix whose entries are independent mean zero subgaussian random variables whose subgaussian moments are bounded by* 1. *Then*

$$\mathbb{P}\big(s_{\max}(A) > C(\sqrt{N} + \sqrt{n}) + t\big) \le 2e^{-ct^2}, \quad t \ge 0.$$

Here and elsewhere in this paper, C, C_1, c, c_1 denote positive absolute constants.

Proof (sketch). We will sketch the proof for $N = n$; the general case is similar. The expression $s_{\max}(A) = \max_{x,y \in S^{n-1}} \langle Ax, y\rangle$ motivates us to first control the random variables $\langle Ax, y\rangle$ individually for each pair of vectors x, y on the unit Euclidean sphere S^{n-1}, and afterwards take the union bound over all such pairs. For fixed $x, y \in S^{n-1}$ the expression $\langle Ax, y\rangle = \sum_{i,j} a_{ij} x_j y_i$ is a sum of independent random variables, where a_{ij} denote the independent entries of A. If a_{ij} were standard normal random variables, the rotation invariance of the Gaussian distribution would imply that $\langle Ax, y\rangle$ is again a standard normal random variable. This property generalizes to subgaussian random variables. Indeed, using moment generating functions one can show that a normalized sum of mean zero subgaussian random variables is again a subgaussian random variable, although the subgaussian moment may increase by an absolute constant factor. Thus

$$\mathbb{P}\big(\langle Ax, y\rangle > s\big) \le 2e^{-cs^2}, \quad s \ge 0.$$

Obviously, we cannot finish the argument by taking the union bound over infinite (even uncountable) number of pairs x, y on the sphere S^{n-1}. In order to reduce the number of such pairs, we discretize S^{n-1} by considering its ε-net $\mathcal{N}_\varepsilon$ in the Euclidean norm, which is a subset of the sphere that approximates every point of the sphere up to error ε. An approximation argument yields

$$s_{\max}(A) = \max_{x,y \in S^{n-1}} \langle Ax, y\rangle \le (1 - \varepsilon)^{-2} \max_{x,y \in \mathcal{N}_\varepsilon} \langle Ax, y\rangle \quad \text{for } \varepsilon \in (0, 1).$$

To gain a control over the size of the net $\mathcal{N}_\varepsilon$, we construct it as a maximal ε-separated subset of S^{n-1}; then the balls with centers in $\mathcal{N}_\varepsilon$ and radii $\varepsilon/2$ form a packing inside the centered ball of radius $1+\varepsilon/2$. A volume comparison gives the useful bound on the cardinality of the net: $|\mathcal{N}_\varepsilon| \le (1+2/\varepsilon)^n$. Choosing for example $\varepsilon = 1/2$, we are well prepared to take the union bound:

$$\mathbb{P}\big(s_{\max}(A) > 4s\big) \le \mathbb{P}\big(\max_{x,y\in\mathcal{N}_\varepsilon} \langle Ax, y\rangle > s\big) \le |\mathcal{N}_\varepsilon| \max_{x,y\in\mathcal{N}_\varepsilon} \mathbb{P}\big(\langle Ax, y\rangle > s\big) \le 5^n \cdot 2e^{-cs^2}.$$

We complete the proof by choosing $s = C\sqrt{n} + t$ with appropriate constant C. □

By integration, one can easily deduce from Proposition 2.4 the correct expectation bound $\mathbb{E}s_{\max}(A) \le C_1(\sqrt{N} + \sqrt{n})$. This latter bound actually holds under much weaker moment assumptions. Similarly to Theorem 2.1, the weakest possible fourth moment assumption suffices here. R. Latala [46] obtained the following general result for matrices with not identically distributed entries:

Theorem 2.5 (Largest singular value: fourth moment, non-iid entries [46]). *Let A be a random matrix whose entries a_{ij} are independent mean zero random variables with finite fourth moment. Then*

$$\mathbb{E}s_{\max}(A) \le C\left[\max_i \left(\sum_j \mathbb{E}a_{ij}^2\right)^{1/2} + \max_j \left(\sum_i \mathbb{E}a_{ij}^2\right)^{1/2} + \left(\sum_{i,j} \mathbb{E}a_{ij}^4\right)^{1/4}\right].$$

For random Gaussian matrices, a much sharper result than in Proposition 2.4 is due to Gordon [31, 32, 33]:

Theorem 2.6 (Exteme singular values of Gaussian matrices, see [19]). *Let A be an $N \times n$ matrix whose entries are independent standard normal random variables. Then*

$$\sqrt{N} - \sqrt{n} \le \mathbb{E}s_{\min}(A) \le \mathbb{E}s_{\max}(A) \le \sqrt{N} + \sqrt{n}.$$

This result is a consequence of the sharp comparison inequalities for Gaussian processes due to Slepian and Gordon, see [31, 32, 33] and [49, Section 3.3].

Tracy-Widom fluctuations One can deduce from Theorem 2.6 a deviation inequality for the extreme singular values. It follows formally by using the concentration of measure in the Gauss space. Since the $s_{\min}(A)$, $s_{\max}(A)$ are 1-Lipschitz functions of A considered as a vector in $\mathbb{R}^{Nn}$, we have

$$\mathbb{P}\big(\sqrt{N} - \sqrt{n} - t \le s_{\min}(A) \le s_{\max}(A) \le \sqrt{N} + \sqrt{n} + t\big) \ge 1 - 2e^{-t^2/2}, \quad t \ge 0, \tag{2.3}$$

see [19]. For general random matrices with independent bounded entries, one can use Talagrand's concentration inequality for convex Lipschitz functions on the cube [76, 77]. Namely, suppose the entries of A are independent, have mean zero, and are uniformly bounded by 1. Since $s_{\max}(A)$ is a convex function of A, Talagrand's concentration inequality implies

$$\mathbb{P}\big(|s_{\max}(A) - \mathrm{Median}(s_{\max}(A))| \geq t\big) \leq 2e^{-t^2/2}.$$

Although the precise value of the median is unknown, integration of the previous inequality shows that $|\mathbb{E}s_{\max}(A) - \mathrm{Median}(s_{\max}(A))| \leq C$. The same deviation inequality holds for symmetric random matrices.

Inequality (2.3) is optimal for large t because $s_{\max}(A)$ is bounded below by the magnitude of every entry of A which has the Gaussian tail. But for small deviations, say for $t < 1$, inequality (2.3) is meaningless. Tracy-Widom law predicts a different tail behavior for small deviations t. It must follow the tail decay of the Tracy–Widom function F_1, which is not subgaussian [3], [39]:

$$c\exp(-C\tau^{3/2}) \leq 1 - F_1(\tau) \leq C\exp(-C'\tau^{3/2}) \quad \tau \geq 0.$$

The concentration of this type for Hermitian complex and real Gaussian matrices (Gaussian Unitary Ensemble and Gaussian Orthogonal Ensemble) was proved by Ledoux [48] and Aubrun [3]. Recently, Feldheim and Sodin [25] introduced a general approach, which allows to prove the asymptotic Tracy–Widom law and its non-asymptotic counterpart at the same time. Moreover, their method is applicable to random matrices with independent subgaussian entries both in symmetric and non-symmetric case. In particular, for an $N \times n$ random matrix A with independent subgaussian entries they proved that

$$p(\tau) := \mathbb{P}\big(s_{\max}(A) \geq \sqrt{N} + \sqrt{n} + \tau\sqrt{N}\big) \leq C\exp(-cn\tau^{3/2}) \quad \tau \geq 0. \tag{2.4}$$

Bounds (2.3) and (2.4) show that the tail behavior of the maximal singular value is essentially different for small and large deviations: $p(\tau)$ decays like $\exp(-cn\tau^{3/2})$ for $\tau \leq c(n/N)^2$ and like $\exp(-c_1 N\tau^2)$ for larger τ. For square matrices the meaning of this phenomenon is especially clear. Large deviations of $s_{\max}(A)$ are produced by bursts of single entries: both $\mathbb{P}(s_{\max}(A) \geq \mathbb{E}s_{\max}(A) + t)$ and $\mathbb{P}(|a_{1,1}| \geq \mathbb{E}s_{\max}(A) + t)$ are of the same order $\exp(-ct^2)$ for $t \geq \mathbb{E}s_{\max}(A)$. In contrast, for small deviations (for smaller t) the situation becomes truly multidimensional, and Tracy-Widom type asymptotics appears.

The method of [25] also addresses the more difficult smallest singular value. For an $N \times n$ random matrix A whose dimensions are not too close to each other Feldheim and Sodin [25] proved the Tracy–Widom law for the smallest singular value together with a non-asymptotic version of the bound $s_{\min}(A) \sim \sqrt{N} - \sqrt{n}$:

$$\mathbb{P}\Big(s_{\min}(A) \leq \sqrt{N} - \sqrt{n} - \tau\sqrt{N}\cdot\frac{N}{N-n}\Big) \leq \frac{C}{1-\sqrt{n/N}}\exp(-c'n\tau^{3/2}). \tag{2.5}$$

3. The Smallest Singular Value

Qualitative invertibility problem In this section we focus on the behavior of the smallest singular value of random $N \times n$ matrices with independent entries. The smallest singular value – the *hard edge* of the spectrum – is generally more difficult and less amenable to analysis by classical methods of random matrix theory than the largest singular value, the "soft edge". The difficulty especially manifests itself for square matrices ($N = n$) or almost square matrices ($N - n = o(n)$). For example, we were guided so far by the asymptotic prediction $s_{\min}(A) \sim \sqrt{N} - \sqrt{n}$, which obviously becomes useless for square matrices.

A remarkable example is provided by $n \times n$ *random Bernoulli matrices* A, whose entries are independent ± 1 valued symmetric random variables. Even the *qualitative invertibility problem*, which asks to estimate the probability that A is invertible, is nontrivial in this situation. Komlós [44, 45] showed that A is invertible asymptotically almost surely: $p_n := \mathbb{P}(s_{\min}(A) = 0) \to 0$ as $n \to \infty$. Later Kahn, Komlos and Szemeredi [43] proved that the singularity probability satisfies $p_n \le c^n$ for some $c \in (0, 1)$. The base c was gradually improved in [78, 81], with the latest record of $p_n = (1/\sqrt{2} + o(1))^n$ obtained in [12]. It is conjectured that the dominant source of singularity of A is the presence of two rows or two columns that are equal up to a sign, which would imply the best possible bound $p_n = (1/2 + o(1))^n$.

Quantitative invertibility problem The previous problem is only concerned with whether the hard edge $s_{\min}(A)$ is zero or not. This says nothing about the *quantitative invertibility problem* of the typical size of $s_{\min}(A)$. The latter question has a long history. Von Neumann and his associates used random matrices as test inputs in algorithms for numerical solution of systems of linear equations. The accuracy of the matrix algorithms, and sometimes their running time as well, depends on the condition number $\kappa(A) = s_{\max}(A)/s_{\min}(A)$. Based on heuristic and experimental evidence, von Neumann and Goldstine predicted that

$$s_{\min}(A) \sim n^{-1/2}, \quad s_{\max}(A) \sim n^{1/2} \quad \text{with high probability} \tag{3.1}$$

which together yield $\kappa(A) \sim n$, see [92, Section 7.8]. In Section 2 we saw several results establishing the second part of (3.1), for the largest singular value.

Estimating the smallest singular value turned out to be more difficult. A more precise form of the prediction $s_{\min}(A) \sim n^{-1/2}$ was repeated by Smale [73] and proved by Edelman [20] and Szarek [79] for *random Gaussian matrices* A, those with i.i.d. standard normal entries. For such matrices, the explicit formula for the joint density of the eigenvalues λ_i of $\frac{1}{n}A^*A$ is available:

$$\mathrm{pdf}(\lambda_1, \dots, \lambda_n) = C_n \prod_{1 \le i < j \le n} |\lambda_i - \lambda_j| \prod_{i=1}^{n} \lambda_i^{-1/2} \exp\left(-\sum_{i=1}^{n} \lambda_i / 2 \right).$$

Integrating out all the eigenvalues except the smallest one, one can in principle compute its distribution. This approach leads to the following asymptotic result:

Theorem 3.1 (Smallest singular value of Gaussian matrices [20]). *Let $A = A_n$ be an $n \times n$ random matrix whose entries are independent standard normal random variables. Then for every fixed $\varepsilon \geq 0$ one has*

$$\mathbb{P}\big(s_{\min}(A) \leq \varepsilon n^{-1/2}\big) \to 1 - \exp(-\varepsilon - \varepsilon^2/2) \quad \text{as } n \to \infty.$$

The limiting probability behaves as $1 - \exp(-\varepsilon - \varepsilon^2/2) \sim \varepsilon$ for small ε. In fact, the following non-asymptotic bound holds for all n:

$$\mathbb{P}\big(s_{\min}(A) \leq \varepsilon n^{-1/2}\big) \leq \varepsilon, \quad \varepsilon \geq 0. \tag{3.2}$$

This follows from the analysis of Edelman [20]; Sankar, Spielman and Teng [68] provided a different geometric proof of estimate (3.2) up to an absolute constant factor and extended it to non-centered Gaussian distributions.

Smallest singular values of general random matrices These methods do not work for general random matrices, especially those with discrete distributions, where rotation invariance and the joint density of eigenvalues are not available. The prediction that $s_{\min}(A) \sim n^{-1/2}$ has been open even for random Bernoulli matrices. Spielman and Teng conjectured in their ICM 2002 talk [75] that estimate (3.2) should hold for the random Bernoulli matrices up to an exponentially small term that accounts for their singularity probability:

$$\mathbb{P}\big(s_{\min}(A) \leq \varepsilon n^{-1/2}\big) \leq \varepsilon + c^n, \quad \varepsilon \geq 0$$

where $c \in (0,1)$ is an absolute constant. The first polynomial bound on $s_{\min}(A)$ for general random matrices was obtained in [63]. Later Spielman-Teng's conjecture was proved in [65] up to a constant factor, and for general random matrices:

Theorem 3.2 (Smallest singular value of square random matrices [65]). *Let A be an $n \times n$ random matrix whose entries are independent and identically distributed subgaussian random variables with zero mean and unit variance. Then*

$$\mathbb{P}\big(s_{\min}(A) \leq \varepsilon n^{-1/2}\big) \leq C\varepsilon + c^n, \quad \varepsilon \geq 0$$

where $C > 0$ and $c \in (0,1)$ depend only on the subgaussian moment of the entries.

This result addresses both qualitative and quantitative aspects of the invertibility problem. Setting $\varepsilon = 0$ we see that A is invertible with probability at least $1 - c^n$. This generaizes the result of Kahn, Komlos and Szemeredi [43] from Bernoulli to all subgaussian matrices. On the other hand, quantitatively, Theorem 3.2 states that $s_{\min}(A) \gtrsim n^{-1/2}$ with high probability for general random matrices. A corresponding non-asymptotic upper bound $s_{\min}(A) \lesssim n^{-1/2}$ also holds [66], so we have $s_{\min}(A) \sim n^{-1/2}$ as in von Neumann-Goldstine's

prediction. Both these bounds, upper and lower, hold with high probability under the weaker fourth moment assumption on the entries [65, 66].

This theory was extended to *rectangular random matrices* of arbitrary dimensions $N \times n$ in [67]. As we know from Section 2, one expects that $s_{\min}(A) \sim \sqrt{N} - \sqrt{n}$. But this would be incorrect for square matrices. To reconcile rectangular and square matrices we make the following correction of our prediction:

$$s_{\min}(A) \sim \sqrt{N} - \sqrt{n-1} \quad \text{with high probability.} \tag{3.3}$$

For square matrices one would have the correct estimate $s_{\min}(A) \sim \sqrt{n} - \sqrt{n-1} \sim n^{-1/2}$. The following result extends Theorem 3.2 to rectangular matrices:

Theorem 3.3 (Smallest singular value of rectangular random matrices [65]). *Let A be an $n \times n$ random matrix whose entries are independent and identically distributed subgaussian random variables with zero mean and unit variance. Then*

$$\mathbb{P}\big(s_{\min}(A) \le \varepsilon(\sqrt{N} - \sqrt{n-1})\big) \le (C\varepsilon)^{N-n+1} + c^N, \quad \varepsilon \ge 0$$

where $C > 0$ and $c \in (0,1)$ depend only on the subgaussian moment of the entries.

This result has been known for a long time for *tall matrices*, whose the aspect ratio $\lambda = n/N$ is bounded by a sufficiently small constant, see [10]. The optimal bound $s_{\min}(A) \ge c\sqrt{N}$ can be proved in this case using an ε-net argument similar to Proposition 2.4. This was extended in [53] to $s_{\min}(A) \ge c_\lambda \sqrt{N}$ for all aspect ratios $\lambda < 1 - c/\log n$. The dependence of c_λ on the aspect ratio λ was improved in [2] for Bernoulli matrices and in [62] for general subgaussian matrices. Feldheim-Sodin's Theorem 2.3 gives precise Tracy-Widom fluctuations of $s_{\min}(A)$ for tall matrices, but becomes useless for almost square matrices (say for $N < n + n^{1/3}$). Theorem 3.3 is an an optimal result (up to absolute constants) which covers matrices with all aspect ratios from tall to square. Non-asymptotic estimate (3.3) was extended to matrices whose entries have finite $(4+\varepsilon)$-th moment in [93].

Universality of the smallest singular values The limiting distribution of $s_{\min}(A)$ turns out to be universal as dimension $n \to \infty$. We already saw a similar universality phenomenon in Theorem 2.3 for genuinely rectangular matrices. For square matrices, the corresponding result was proved by Tao and Vu [87]:

Theorem 3.4 (Smallest singular value of square matrices: universality [87]). *Let A be an $n \times n$ random matrix whose entries are independent and identically distributed random variables with zero mean, unit variance, and finite*

K-th moment where K is a sufficiently large absolute constant. Let G be an $n \times n$ random matrix whose entries are independent standard normal random variables. Then

$$\mathbb{P}(\sqrt{n} s_{\min}(G) \le t - n^{-c}) - n^c \le \mathbb{P}(\sqrt{n} s_{\min}(A) \le t) \le \mathbb{P}(\sqrt{n} s_{\min}(G) \le t + n^{-c}) + n^c$$

where $c > 0$ depends only on the K-th moment of the entries.

On a methodological level, this result may be compared in classical probability theory to Berry-Esseen theorem (1.1) which establishes polynomial deviations from the limiting distribution, while Theorems 3.2 and 3.3 bear a similarity with large deviation results like (1.2) which give exponentially small tail probabilities.

Sparsity and invertibility: a geometric proof of Theorem 3.2 We will now sketch the proof of Theorem 3.2 given in [65]. This argument is mostly based on geometric ideas, and it may be useful beyond spectral analysis of random matrices.

Looking at $s_{\min}(A) = \min_{x \in S^{n-1}} \|Ax\|_2$ we see that our goal is to bound below $\|Ax\|_2$ uniformly for all unit vectors x. We will do this separately for *sparse vectors* and for *spread vectors* with two very different arguments. Choosing a small absolute constant $c_0 > 0$, we first consider the class of sparse vectors

$$\mathit{Sparse} := \{x \in S^{n-1} : \ |\operatorname{supp}(x)| \le c_0 n\}$$

Establishing invertibility of A on this class is relatively easy. Indeed, when we look at $\|Ax\|_2$ for sparse vectors x of fixed support $\operatorname{supp}(x) = I$ of size $|I| = c_0 n$, we are effectively dealing with the $n \times c_0 n$ submatrix A_I that consists of the columns of A indexed by I. The matrix A_I is tall, so as we said below Theorem 3.3, its smallest singular value can be estimated using the standard ε-net argument. This gives $s_{\min}(A_I) \ge cn^{1/2}$ with probability at least $1 - 2e^{-n}$. This allows us to further take the union bound over $\binom{n}{c_0 n} \le e^{n/2}$ choices of support I, and conclude that with probability at least $1 - 2e^{-n/2}$ we have invertibility on all sparse vectors:

$$\min_{x \in \mathit{Sparse}} \|Ax\|_2 = \min_{|I| \le c_0 n} s_{\min}(A_I) \ge cn^{1/2}. \tag{3.4}$$

We thus obtained a much stronger bound than we need, $n^{1/2}$ instead of $n^{-1/2}$.

Establishing invertibility of A on non-sparse vectors is more difficult because there are too many of them. For example, there are exponentially many vectors on S^{n-1} whose coordinates all equal $\pm n^{-1/2}$ and which have at least a constant distance from each other. This gives us no hope to control such vectors using ε-nets, as any nontrivial net must have cardinality at least 2^n. So let us now focus on this most difficult class of extremely non-sparse vectors

$$\mathit{Spread} := \{x \in S^{n-1} : \ |x_i| \ge c_1 n^{-1/2} \text{ for all } i\}.$$

Once we prove invertibility of A on these spread vectors, the argument can be completed for all vectors in S^{n-1} by an approximation argument. Loosely speaking, if x is close to *Sparse* we can treat x as sparse, otherwise x must have at least cn coordinates of magnitude $|x_i| = O(n^{-1/2})$, which allows us to treat x as spread.

An obvious advantage of spread vectors is that we know the magnitude of all their coefficients. This motivates the following geometric invertibility argument. If A performs extremely poor so that $s_{\min}(A) = 0$, then one of the columns X_k of A lies in the span $H_k = \operatorname{span}(X_i)_{i \neq k}$ of the others. This simple observation can be transformed into a quantitative argument. Suppose $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is a spread vector. Then, for every $k = 1, \dots, n$, we have

$$\begin{aligned} \|Ax\|_2 \geq \operatorname{dist}(Ax, H_k) &= \operatorname{dist}\left(\sum_{i=1}^n x_i X_i, H_k\right) = \operatorname{dist}(x_k X_k, H_k) \\ &= |x_k| \cdot \operatorname{dist}(X_k, H_k) \geq c_1 n^{-1/2} \operatorname{dist}(X_k, H_k). \end{aligned} \tag{3.5}$$

Since the right hand side does not depend on x, we have proved that

$$\min_{x \in Spread} \|Ax\|_2 \geq c_1 n^{-1/2} \operatorname{dist}(X_n, H_n). \tag{3.6}$$

This reduces our task to the geometric problem of independent interest – estimate *the distance between a random vector and an independent random hyperplane*. The expectation estimate $1 \leq \mathbb{E} \operatorname{dist}(X_n, H_n)^2 = O(1)$ follows easily by independence and moment assumptions. But we need a lower bound with high probability, which is far from trivial. This will make a separate story connected to the Littlewood-Offord theory of *small ball probabilities*, which we discuss in Section 4. In particular we will prove in Corollary 4.4 the optimal estimate

$$\mathbb{P}(\operatorname{dist}(X_n, H_n) \leq \varepsilon) \leq C\varepsilon + c^n, \quad \varepsilon \geq 0, \tag{3.7}$$

which is simple for the Gaussian distribution (by rotation invariance) and difficult to prove e.g. for the Bernoulli distribution. Together with (3.6) this means that we proved invertibility on all spread vectors:

$$\mathbb{P}\left(\min_{x \in Spread} \|Ax\|_2 \leq \varepsilon n^{-1/2}\right) \leq C\varepsilon + c^n, \quad \varepsilon \geq 0.$$

This is exactly the type of probability bound claimed in Theorem 3.2. As we said, we can finish the proof by combining with the (much better) invertibility on sparse vectors in (3.4), and by an approximation argument.

4. Littlewood-Offord Theory

Small ball probabilities and additive structure We encountered the following geometric problem in the previous section: *estimate the distance between*

a random vector X with independent coordinates and an independent random hyperplane H in $\mathbb{R}^n$. We need a lower bound on this distance with high probability. Let us condition on the hyperplane H and let $a \in \mathbb{R}^n$ denote its unit normal vector. Writing in coordinates $a = (a_1, \dots, a_n)$ and $X = (\xi_1, \dots, \xi_n)$, we see that

$$\operatorname{dist}(X, H) = \langle a, X \rangle = \left| \sum_{i=1}^{n} a_i \xi_i \right|. \tag{4.1}$$

We need to understand the distribution of sums of independent random variables

$$S = \sum_{i=1}^{n} a_i \xi_i, \quad \|a\|_2 = 1,$$

where $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ is a given coefficient vector, and $\xi_1, \dots, \xi_n$ are independent identically distributed random variables with zero mean and unit variance.

Sums of independent random variables is a classical theme in probability theory. The well-developed area of large deviation inequalities like (1.2) demonstrates that S nicely concentrates around its mean. But our problem is opposite as we need to show that S is *not too concentrated* around its mean 0, and perhaps more generally around any real number. Several results in probability theory starting from the works of Lévy [50], Kolmogorov [42] and Esséen [24] were concerned with the spread of sums of independent random variables, which is quantified as follows:

Definition 4.1. The *Lévy concentration function* of a random variable S is

$$\mathcal{L}(S, \varepsilon) = \sup_{v \in \mathbb{R}} \mathbb{P}(|S - v| \le \varepsilon), \quad \varepsilon \ge 0.$$

Lévy concentration function measures the *small ball probability* [51], the likelihood that S enters a small interval. For continuous distributions one can show that $\mathcal{L}(S, \varepsilon) \lesssim \varepsilon$ for all $\varepsilon \ge 0$. For discrete distributions this may be false. Instead, a new phenomenon arises for discrete distributions which is unseen in large deviation theory: Lévy concentration function depends on the *additive structure* of the coefficient vector a. This is best illustrated on the example where ξ_i are independent Bernoulli random variables (± 1 valued and symmetric). For sparse vectors like $a = 2^{-1/2}(1, 1, 0, \dots, 0)$, Lévy concentration function can be large: $\mathcal{L}(S, 0) = 1/2$. For spread vectors, Berry-Esseen's theorem (1.1) yields a better bound:

$$\text{For } a' = n^{-1/2}(1, 1, \dots, 1), \quad \mathcal{L}(S, \varepsilon) \le C(\varepsilon + n^{-1/2}). \tag{4.2}$$

The threshold $n^{-1/2}$ comes from many cancelations in the sums $\sum \pm 1$ which occur because all coefficients a_i are equal. For less structured a, fewer cancelations occur:

$$\text{For } a'' = n^{-1/2}\left(1 + \frac{1}{n}, 1 + \frac{2}{n}, \dots, 1 + \frac{n}{n}\right), \quad \mathcal{L}(S, 0) \sim n^{-3/2}. \tag{4.3}$$

Studying the influence of additive structure of the coefficient vector a on the spread of $S = \sum a_i \xi_i$ became known as the *Littlewood-Offord problem*. It was initially developed by Littlewood and Offord [52], Erdös and Moser [21, 22], Sárkozy and Szemerédi [69], Halasz [40], Frankl and Füredi [26]. For example, if all $|a_i| \geq 1$ then $\mathcal{L}(S, 1) \leq Cn^{-1/2}$ [52, 21], which agrees with (4.2). Similarly, a general fact behind (4.3) is that if $|a_i - a_j| \geq 1$ for all $i \neq j$ then $\mathcal{L}(S, 1) \leq Cn^{-3/2}$ [22, 69, 40].

New results on Lévy concentration function Problems of invertibility of random matrices motivated a recent revisiting of LO problem by Tao and Vu [83, 84, 86, 88], the authors [65, 67], Friedland and Sodin [27]. Additive structure of the coefficient vector a is related to the shortest arithmetic progression into which it embeds. This length is conveniently expressed as the *least common denominator* $\operatorname{lcd}(a)$ defined as the smallest $\theta > 0$ such that $\theta a \in \mathbb{Z}^n \setminus 0$. Examples suggest that Lévy concentration function should be inversely proportional to the least common denominator: $\operatorname{lcd}(a') = n^{1/2} \sim 1/\mathcal{L}(S, 0)$ in (4.2) and $\operatorname{lcd}(a'') = n^{3/2} \sim 1/\mathcal{L}(S, 0)$ in (4.3). This is not a coincidence. But to state a general result, we will need to consider a more stable version of the least common denominator. Given an accuracy level $\alpha > 0$, we define the *essential least common denominator*

$$\operatorname{lcd}_\alpha(a) := \inf \left\{ \theta > 0 : \ \operatorname{dist}(\theta a, \mathbb{Z}^n) \leq \min \left(\frac{1}{10} \|\theta a\|_2, \alpha \right) \right\}.$$

The requirement $\operatorname{dist}(\theta a, \mathbb{Z}^n) \leq \frac{1}{10} \|\theta a\|_2$ ensures approximation of θa by non-trivial integer points, those in a non-trivial cone in the direction of a. The constant $\frac{1}{10}$ is arbitrary and it can be replaced by any other constant in $(0, 1)$. One typically uses this concept for accuracy levels $\alpha = c\sqrt{n}$ with a small constant c such as $c = \frac{1}{10}$. The inequality $\operatorname{dist}(\theta a, \mathbb{Z}^n) \leq \alpha$ yields that most of the coordinates of θa are within a small constant distance from integers. For such α, in examples (4.2) and (4.3) one has as before $\operatorname{lcd}_\alpha(a') \sim n^{1/2}$ and $\operatorname{lcd}_\alpha(a'') \sim n^{3/2}$. Here we state and sketch a proof of a general Littlewood-Offord type result from [67].

Theorem 4.2 (Lévy concentration function via additive structure). *Let $\xi_1, \dots, \xi_n$ be independent identically distributed mean zero random variables, which are well spread: $p := \mathcal{L}(\xi_k, 1) < 1$. Then, for every coefficient vector $a = (a_1, \dots, a_n) \in S^{n-1}$ and every accuracy level $\alpha > 0$, the sum $S = \sum_{i=1}^n a_i \xi_i$ satisfies*

$$\mathcal{L}(S, \varepsilon) \leq C\varepsilon + C/\operatorname{lcd}_\alpha(a) + Ce^{-c\alpha^2}, \quad \varepsilon \geq 0, \tag{4.4}$$

where $C, c > 0$ depend only on the spread p.

Proof. A classical Esseen's concentration inequality [24] bounds the Lévy concentration function of an arbitrary random variable Z by the L_1 norm of its

characteristic function $\phi_Z(\theta) = \mathbb{E}\exp(i\theta Z)$ as follows:

$$\mathcal{L}(Z,1) \le C\int_{-1}^{1} |\phi_Z(\theta)|\,d\theta. \tag{4.5}$$

One can prove this inequality using Fourier inversion formula, see [80, Section 7.3].

We will show how to prove Theorem 4.2 for Bernoulli random variables ξ_i; the general case requires an additional argument. Without loss of generality we can assume that $\mathrm{lcd}_\alpha(a) \ge \frac{1}{\pi\varepsilon}$. Applying (4.5) for $Z = S/\varepsilon$, we obtain by independence that

$$\mathcal{L}(S,\varepsilon) \le C\int_{-1}^{1} |\phi_S(\theta/\varepsilon)|\,d\theta = C\int_{-1}^{1}\prod_{j=1}^{n} |\phi_j(\theta/\varepsilon)|\,d\theta,$$

where $\phi_j(t) = \mathbb{E}\exp(ia_j\xi_j t) = \cos(a_j t)$. The inequality $|x| \le \exp(-\frac{1}{2}(1-x^2))$ which is valid for all $x \in \mathbb{R}$ implies that

$$|\phi_j(t)| \le \exp\left(-\frac{1}{2}\sin^2(a_j t)\right) \le \exp\left(-\frac{1}{2}\,\mathrm{dist}\left(\frac{a_j t}{\pi},\mathbb{Z}\right)^2\right).$$

Therefore

$$\mathcal{L}(S,\varepsilon) \le C\int_{-1}^{1}\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\mathrm{dist}\left(\frac{a_j\theta}{\pi\varepsilon},\mathbb{Z}\right)^2\right)d\theta = C\int_{-1}^{1}\exp\Big(-\frac{1}{2}f^2(\theta)\Big)\,d\theta \tag{4.6}$$

where $f(\theta) = \mathrm{dist}\big(\frac{\theta}{\pi\varepsilon}a, \mathbb{Z}^n\big)$. Since $\mathrm{lcd}_\alpha(a) \ge \frac{1}{\pi\varepsilon}$, the definition of the essential least common denominator implies that for every $\theta \in [-1,1]$ we have $f(\theta) \ge \min(\frac{\theta}{10\pi\varepsilon}\|a\|_2, \alpha)$. Since by assumption $\|a\|_2 = 1$, it follows that

$$\exp\left(-\frac{1}{2}f^2(\theta)\right) \le \exp\left(-\frac{1}{2}\left(\frac{\theta}{10\pi\varepsilon}\right)^2\right) + \exp(-\alpha^2/2).$$

Substituting this into (4.6) yields $\mathcal{L}(S,\varepsilon) \le C_1(\varepsilon + 2\exp(-\alpha^2/2))$ as required. □

Theorem 4.2 justifies our empirical observation that Lévy concentration function is proportional to the amount of structure in the coefficient vector, which is measured by the (reciprocal of) its essential least common denominator. As we said, this result is typically used for accuracy level $\alpha = c\sqrt{n}$ with some small positive constant c. In this case, the term $Ce^{-c\alpha^2}$ in (4.4) is exponentially small in n (thus negligible in applications), and the term $C\varepsilon$ is optimal for continuous distributions.

Theorem 4.2 performs best for totally unstructured coefficient vectors a, those with exponentially large $\mathrm{lcd}_\alpha(a)$. Heuristically, this should be the case for random vectors, as randomness should destroy any structure. While this is not true for general vectors with independent coordinates (e.g. for equal coordinates with random signs), it is true for *normals of random hyperplanes*:

Theorem 4.3 (Random vectors are unstructured [65]). *Let X_i be random vectors in $\mathbb{R}^n$ whose coordinates are independent and identically distributed subgaussian random variables with zero mean and unit variance. Let $a \in \mathbb{R}^n$ denote a unit normal vector of $H = \mathrm{span}(X_1, \dots, X_{n-1})$. Then, with probability at least $1 - e^{-cn}$,*

$$\mathrm{lcd}_\alpha(a) \geq e^{cn} \quad \text{for } \alpha = c\sqrt{n},$$

where $c > 0$ depends only on the subgaussian moment.

Therefore for random normals a, Theorem 4.2 yealds with high probability a very strong bound on Lévy concentration function:

$$\mathcal{L}(S, \varepsilon) \leq C\varepsilon + c^n, \quad \varepsilon \geq 0. \tag{4.7}$$

This brings us back to the distance problem considered in the beginning of this section, which motivated our study of Lévy concentration function:

Corollary 4.4 (Distance between random vectors and hyperplanes [65]). *Let X_i be random vectors as in Theorem 4.3, and $H_n = \mathrm{span}(X_1, \dots, X_{n-1})$. Then*

$$\mathbb{P}\big(\operatorname{dist}(X_n, H_n) \leq \varepsilon\big) \leq C\varepsilon + c^n, \quad \varepsilon \geq 0,$$

where $C, c > 0$ depend only on the subgaussian moment.

Proof. As was noticed in (4.1), we can write $\operatorname{dist}(X_n, H_n)$ as a sum of independent random variables, and then bound it using (4.7). □

Corollary 4.4 offers us exactly the missing piece (3.7) in our proof of the invertibility Theorem 3.2. This completes our analysis of invertibility of square matrices.

Remark. These methods generalize to rectangular matrices [67, 93]. For example, Corollary 4.4 can be extended to compute the distance between random vectors and subspaces of arbitrary dimension [67]: for $H_n = \mathrm{span}(X_1, \dots, X_{n-d})$ we have $(\mathbb{E} \operatorname{dist}(X_n, H_n)^2)^{1/2} = \sqrt{d}$ and

$$\mathbb{P}\big(\operatorname{dist}(X_n, H_n) \leq \varepsilon\sqrt{d}\big) \leq (C\varepsilon)^d + c^n, \quad \varepsilon \geq 0.$$

5. Applications

The applications of non-asymptotic theory of random matrices are numerous, and we cannot cover all of them in this note. Instead we concentrate on three different results pertaining to the classical random matrix theory (Circular Law), signal processing (compressed sensing), and geometric functional analysis and theoretical computer science (short Khinchin's inequality and Kashin's subspaces).

Circular law Asymptotic theory of random matrices provides an important source of heuristics for non-asymptotic results. We have seen an illustration of this in the analysis of the extreme singular values. This interaction between the asymptotic and non-asymptotic theories goes the other way as well, as good non-asymptotic bounds are sometimes crucial in proving the limit laws. One remarkable example of this is the circular law which we will discuss now.

Consider a family of $n \times n$ matrices A whose entries are independent copies of a random variable X with mean zero and unit variance. Let μ_n be the empirical measure of the eigenvalues of the matrix $B_n = \frac{1}{\sqrt{n}} A_n$, i.e. the Borel probability measure on $\mathbb{C}$ such that $\mu_n(E)$ is the fraction of the eigenvalues of B_n contained in E. A long-standing conjecture in random matrix theory, which is called the circular law, suggested that *the measures μ_n converge to the normalized Lebesgue measure on the unit disc*. The convergence here can be understood in the same sense as in the Wigner's semicircle law. The circular law was originally proved by Mehta [56] for random matrices with standard normal entries. The argument used the explicit formula for joint density of the eigenvalues, so it could not be extended to other classes of random matrices. While the formulation of Wigner's semicircle law and the circular law look similar, the methods used to prove the former are not applicable to the latter. The reason is that the spectrum of a general matrix, unlike that of a Hermitian matrix, is unstable: a small change of the entries may cause a significant change of the spectrum (see [6]). Girko [30] introduced a new approach to the circular law based on considering the real part of the Stieltjes transform of measures μ_n. For $z = x + iy$ the real Stieltjes transform is defined by the formula

$$S_{nr}(z) = \mathrm{Re}\left(\frac{1}{n}\mathrm{Tr}(B_n - zI_n)^{-1}\right) = -\frac{\partial}{\partial x}\left(\frac{1}{n}\log|\det(B_n - zI)|\right).$$

Since $|\det(B_n - zI)|^2 = \det(B_n - zI)(B_n - zI)^*$, this is the same as

$$S_{nr}(z) = -\frac{1}{2}\frac{\partial}{\partial x}\left(\frac{1}{n}\log|\det(B_n - zI)(B_n - zI)^*|\right) = -\frac{1}{2}\frac{\partial}{\partial x}\left(\frac{1}{n}\sum_{j=1}^{n}\log s_j^{(n)}(z)\right),$$

where $s_1^{(n)}(z) \geq \ldots \geq s_n^{(n)}(z) \geq 0$ are the eigenvalues of the *Hermitian* matrix $(B_n - zI)(B_n - zI)^*$, or in other words, the squares of the singular values of the matrix $V_n = B_n - zI$. Girko's argument reduces the proof of the circular law to the convergence of real Stieltjes transforms, and thus to the behavior of the sum above. The logarithmic function is unbounded at 0 and ∞. To control the behavior near ∞, one has to use the bound for the largest singular value of V_n, which is relatively easy. The analysis of the behavior near 0 requires bounds on the smallest singular value of V_n, and is therefore more difficult.

Girko's approach was implemented by Bai [4], who proved the circular law for random matrices whose entries have bounded sixth moment and bounded

density. The bounded density condition was sufficient to take care of the smallest singular value problem. This result was the first manifestation of the universality of the circular law. Still, it did not cover some important classes of random matrices, in particular random Bernoulli matrices. The recent results on the smallest singular value led to a significant progress on establishing the universality of the circular law. A crucial step was done by Götze and Tikhomirov [34] who extended the circular law to all subgaussian matrices using [63]. In fact, the results of [34] actually extended it to all random entries with bounded fourth moment. This was further extended to random variables having bounded moment $2+\varepsilon$ in [35, 82]. Finally, in [85] Tao and Vu proved the Circular Law in full generality, with no assumptions besides the unit variance. Their approach was based on the smallest singular value bound from [82] and a novel *replacement principle* which allowed them to treat the other singular values.

Compressed Sensing Non-asymptotic random matrix theory provides a right context for the analysis of random measurements in the newly developed area of compressed sensing, see the ICM 2006 talk of Candes [14]. Compressed sensing is an area of information theory and signal processing which studies efficient techniques to reconstruct a signal from a small number of measurements by utilizing the prior knowledge that the signal is sparse [18].

Mathematically, one seeks to reconstruct an unknown signal $x \in \mathbb{R}^n$ from some m linear measurements viewed as a vector $Ax \in \mathbb{R}^m$, where A is some known $m \times n$ matrix called the *measurement matrix*. In the interesting case $m < n$, the problem is underdetermined and we are interested in the sparsest solution:

$$\text{minimize } \|x^*\|_0 \text{ subject to } Ax^* = Ax, \tag{5.1}$$

where $\|x\|_0 = |\operatorname{supp}(x)|$. This optimization problem is highly non-convex and computationally intractable. So one considers the following convex relaxation of (5.1), which can be efficiently solved by convex programming methods:

$$\text{minimize } \|x^*\|_1 \text{ subject to } Ax^* = Ax, \tag{5.2}$$

where $\|x\|_1 = \sum_{i=1}^n |x_i|$ denotes the ℓ_1 norm.

One would then need to find conditions when problems (5.1) and (5.2) are equivalent. Candes and Tao [16] showed that this occurs when the measurement matrix A is a *restricted isometry*. For an integer $s \le n$, the restricted isometry constant $\delta_s(A)$ is the smallest number $\delta \ge 0$ which satisfies

$$(1-\delta)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta)\|x\|_2^2 \quad \text{for all } x \in \mathbb{R}^n,\ |\operatorname{supp}(x)| \le s. \tag{5.3}$$

Geometrically, the restricted isometry property guarantees that the geometry of s-sparse vectors x is well preserved by the measurement matrix A. In turns

out that in this situation one can reconstruct x from Ax by the convex program (5.2):

Theorem 5.1 (Sparse reconstruction using convex programming [16]). *Assume $\delta_{2s} \le c$. Then the solution of (5.2) equals x whenever $|\operatorname{supp}(x)| \le s$.*

A proof with $c = \sqrt{2}-1$ is given in [15]; the current record is $c = 0.472$ [13].

Restricted isometry property can be interpreted in terms of the *extreme singular values of submatrices* of A. Indeed, (5.3) equivalently states that the inequality

$$\sqrt{1-\delta} \le s_{\min}(A_I) \le s_{\max}(A_I) \le \sqrt{1+\delta}$$

holds for all $m \times s$ submatrices A_I, those formed by the columns of A indexed by sets I of size s. In light of Sections 2 and 3, it is not surprising that the best known restricted isometry matrices are *random matrices*. It is actually an open problem to construct *deterministic* restricted isometry matrices as in Theorem 5.2 below.

The following three types of random matrices are extensively used as measurement matrices in compressed sensing: *Gaussian, Bernoulli, and Fourier*. Here we summarize their restricted isometry properties, which have the common remarkable feature: *the required number of measurements m is roughly proportional to the sparsity level s* rather than the (possibly much larger) dimension n.

Theorem 5.2 (Random matrices are restricted isometries). *Let m, n, s be positive integers, $\varepsilon, \delta \in (0,1)$, and let A be an $m \times n$ measurement matrix.*

1. Suppose the entries of A are independent and identically distributed subgaussian random variables with zero mean and unit variance. Assume that

$$m \ge Cs \log(2n/s)$$

where C depends only on ε, δ, and the subgaussian moment. Then with probability at least $1-\varepsilon$, the matrix $\bar{A} = \frac{1}{\sqrt{m}} A$ is a restricted isometry with $\delta_s(\bar{A}) \le \delta$.

2. Let A be a random Fourier matrix obtained from the $n \times n$ discrete Fourier transform matrix by choosing m rows independently and uniformly. Assume that

$$m \ge Cs \log^4(2n). \tag{5.4}$$

where C depends only on ε and δ. Then with probability at least $1-\varepsilon$, the matrix $\bar{A} = \frac{1}{\sqrt{n}} A$ is a restricted isometry with $\delta_s(\bar{A}) \le \delta$.

For random subgaussian matrices this result was proved in [9, 57] by an ε-net argument, where one first checks the deviation inequality $|\|Ax\|_2^2 - 1| \le \delta$ with exponentially high probability for a fixed vector x as in (5.3), and afterwards lets x run over some fine net. For random Fourier matrices the problem is harder. It was first addressed in [17] with a little higher exponent than in (5.4); the exponent 4 was obtained in [64], and it is conjectured that the optimal exponent is 1.

Short Khinchin's inequality and Kashin's subspaces Let $1 \le p < \infty$. The classical Khinchin's inequality states that there exist constants A_p, B_p such that for all $x = (x_1, \dots, x_n) \in \mathbb{R}^n$

$$A_p \|x\|_2 \le \left(\operatorname*{Ave}_{\varepsilon \in \{-1,1\}^n} \left| \sum_{j=1}^n \varepsilon_j x_j \right|^p \right)^{1/p} \le B_p \|x\|_2.$$

The average here is taken over all 2^n possible choices of signs ε (it is the same as the expectation with respect to independent Bernoulli random variables ε_j). Since the mid-seventies, the question was around whether Khinchin's inequality holds for averages over some small sets of signs ε. A trivial lower bound follows by a dimension argument: such a set must contain at least n points. Here we shall discuss only the case $p = 1$, which is of considerable interest for computer science. This problem can be stated more precisely as follows: as follows:

> Given $\delta > 0$, find $\alpha(\delta), \beta(\delta) > 0$ and construct a set $V \subset \{-1, 1\}^n$ of cardinality less than $(1+\delta)n$ such that for all $x = (x_1, \dots, x_n) \in \mathbb{R}^n$
>
> $$\alpha(\delta)\|x\|_2 \le \operatorname*{Ave}_{\varepsilon \in V} \left| \sum_{j=1}^n \varepsilon_j x_j \right| \le \beta(\delta)\|x\|_2. \tag{5.5}$$

The first result in this direction belongs to Schechtman [70] who found an affirmative solution to this problem for δ greater than some absolute constant. He considered a set V consisting of $N = \lfloor (1+\delta)n \rfloor$ independent random ± 1 vectors, which can be written as an $N \times n$ random Bernoulli matrix A. In the matrix language, the inequality above reads $\alpha(\delta)\|x\|_2 \le N^{-1}\|Ax\|_1 \le \beta(\delta)\|x\|_2$ for all $x \in \mathbb{R}^n$. This means that one can take

$$\alpha(\delta) = N^{-1} \inf_{x \in S^{n-1}} \|Ax\|_1, \quad \beta(\delta) = N^{-1} \sup_{x \in S^{n-1}} \|Ax\|_1.$$

These expressions bear a similarity to the smallest and the largest singular values of the matrix A. In fact, up to the coefficient N^{-1}, $\beta(\delta)$ is the norm of A considered as a linear operator from ℓ_2^n to ℓ_1^n, and $\alpha(\delta)$ is the reciprocal of the norm of its inverse. Schechtman's theorem can now be derived using the ε-net argument.

The case of small δ is more delicate. For a random A, the bound for $\beta(\delta) \le C$ can be obtained by the ε-net argument as before. However, an attempt to apply this argument for $\alpha(\delta)$ runs into to the same problems as for the smallest singular value. For any fixed $\delta > 0$ the solution was first obtained first by Johnson and Schechtman [38] who showed that there *exists* V satisfying (5.5) with $\alpha(\delta) = c^{1/\delta}$. In [54] this was established for a random set V (or

a random matrix A) with the same bound on $\alpha(\delta)$. Furthermore, the result remains valid even when δ depends on n, as long as $\delta \geq c/\log n$. The proof uses the smallest singular value bound from [53] in a crucial way. The bound on $\alpha(\delta)$ has been further improved in [2], also using the singular value approach. Finally, a theorem in [62] asserts that for a random set V the inequalities (5.5) hold with high probability for

$$\alpha(\delta) = c\delta^2, \quad \beta(\delta) = C.$$

Moreover, the result holds for all $\delta > 0$ and n, without any restrictions. The proof combines the methods of [63] and a geometric argument based on the structure of a section of the ℓ_1^n ball. The probability estimate of [62] can be further improved if one replaces the small ball probability bound of [63] with that of [65].

The short Khinchin inequality shows also that the ℓ_1 and ℓ_2 norms are equivalent on a random subspace $E := A\mathbb{R}^n \subset \mathbb{R}^N$. Indeed, if A is an $N \times n$ random matrix, then with high probability every vector $x \in \mathbb{R}^n$ satisfies $\alpha(\delta)\|x\|_2 \leq N^{-1}\|Ax\|_1 \leq N^{-1/2}\|Ax\|_2 \leq C\|x\|_2$. The second inequality here is Cauchy-Schwartz, and the third one is the largest singular value bound. Thierefore

$$C^{-1}\alpha(\delta)\|y\|_2 \leq N^{-1/2}\|y\|_1 \leq \|y\|_2 \qquad \text{for all } y \in E. \tag{5.6}$$

Subspaces E possessing property (5.6) are called *Kashin's subspaces*. The classical Dvoretzky theorem states that a high-dimensional Banach space has a subspace which is close to Euclidean [59]. The dimension of such subspace depends on the geometry of the ambient space. Milman proved that such subspaces always exist in dimension $c\log n$, where n is the dimension of the ambient space [58] (see also [59]). For the space ℓ_1^n the situation is much better, and such subspaces exist in dimension $(1-\delta)n$ for any constant $\delta > 0$. This was first proved by Kashin [41] also using a random matrix argument. Obviously, as $\delta \to 0$, the distance between the ℓ_1 and ℓ_2 norms on such subspace grows to ∞. The optimal bound for this distance has been found by Garnaev and Gluskin [28] who used subspaces generated by Gaussian random matrices.

Kashin's subspaces turned out to be useful in theoretical computer science, in particular in the nearest neighbor search [36] and in compressed sensing. At present no deterministic construction is known of such subspaces of dimension n proportional to N. The result of [62] shows that a $\lfloor(1+\delta)n\rfloor \times n$ random Bernoulli matrix defines a Kashin's subspace with $\alpha(\delta) = c\delta^2$. A random Bernoulli matrix is computationally easier to implement than a random Gaussian matrix, while the distance between the norms is not much worse than in the optimal case. At the same time, since the subspaces generated by a Bernoulli matrix are spanned by random vertices of the discrete cube, they have relatively simple structure, which is possible to analyze.

References

[1] R. Adamczak, A. Litvak, A. Pajor, N. Tomczak-Jaegermann, *Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles*, J. Amer. Math. Soc. 23 (2010), 535–561.

[2] S. Artstein-Avidan, O. Friedland, V. D. Milman, S. Sodin, *Polynomial bounds for large Bernoulli sections of l_1^N*, Israel J. Math. 156 (2006), 141–155.

[3] G. Aubrun, *A sharp small deviation inequality for the largest eigenvalue of a random matrix*, Séminaire de Probabilités XXXVIII, 320–337, Lecture Notes in Math., 1857, Springer, Berlin, 2005.

[4] Z. D. Bai, *Circular law*, Ann. Probab. 25 (1997), no. 1, 494–529.

[5] Z. D. Bai, *Methodologies in spectral analysis of large dimensional random matrices, a review*, Statistica Sinica 9 (1999), 611–677

[6] Z. D. Bai, J. Silverstein, *Spectral analysis of large dimensional random matrices*, 2nd ed., Springer Series in Statistics, Springer, New York, 2010.

[7] Z. D. Bai, J. Silverstein, Y. Q. Yin, *A note on the largest eigenvalue of a large-dimensional sample covariance matrix*, J. Multivariate Anal. 26 (1988), 166–168.

[8] Z. D. Bai,Y. Q. Yin, *Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix*, Ann. Probab. 21 (1993), 1275–1294.

[9] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx. 28 (2008), 253–263.

[10] G. Bennett, L. E. Dor, V. Goodman, W. B. Johnson, C. M. Newman, *On uncomplemented subspaces of L_p, $1 < p < 2$*, Israel J. Math. 26 (1977), 178–187.

[11] A. C. Berry, *The accuracy of the Gaussian approximation to the sum of independent variables*, Trans. Amer. Math. Soc., 49 (1941), 122–136.

[12] J. Bourgain, P. Wood, V. Vu, *On the singularity probability of random discrete matrices*, submitted.

[13] T. Cai, L. Wang, and G. Xu, *Shifting Inequality and Recovery of Sparse Signals*, IEEE Transactions on Signal Processing, to appear.

[14] E. Candés, *Compressive sampling*, International Congress of Mathematicians. Vol. III, 1433–1452, Eur. Math. Soc., Zürich, 2006.

[15] E. Candés, *The restricted isometry property and its implications for compressed sensing*, C. R. Math. Acad. Sci. Paris 346 (2008), 589–592.

[16] E. Candés, T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory 51 (2005), 4203–4215.

[17] E. Candés, T. Tao, *Near-optimal signal recovery from random projections: universal encoding strategies?* IEEE Trans. Inform. Theory 52 (2006), 5406–5425.

[18] E. Candés, M. B. Wakin, *An Introduction To Compressive Sampling*, IEEE Signal Processing Magazine, V.21, March 2008.

[19] K. R. Davidson, S. J. Szarek, *Local operator theory, random matrices and Banach spaces.* Handbook of the geometry of Banach spaces, Vol. I, 317–366, North-Holland, Amsterdam, 2001.

[20] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl. 9 (1988), 543–560

[21] P. Erdös, *On a lemma of Littlewood and Offord*, Bull. Amer. Math. Soc. 51 (1945), 898–902

[22] P. Erdös, *Extremal problems in number theory*, 1965 Proc. Sympos. Pure Math., Vol. VIII, pp.181–189 AMS, Providence, R.I.

[23] C. G. Esseen, *Fourier analysis of distribution functions. A mathematical study of the laplace Gaussian law*, Acta Math. 77 (1945), 1–125.

[24] C. G. Esseen, *On the Kolmogorov-Rogozin inequality for the concentration function*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 5 (1966), 210–216.

[25] O. Feldheim, S. Sodin, *A universality result for the smallest eigenvalues of certain sample covariance matrices*, Geometric and Functional Analysis, to appear

[26] P. Frankl, Z. Füredi, *Solution of the Littlewood-Offord problem in high dimensions*, Ann. of Math. (2) 128 (1988), 259–270.

[27] O. Friedland, S. Sodin, *Bounds on the concentration function in terms of the Diophantine approximation*, C. R. Math. Acad. Sci. Paris 345 (2007), 513–518.

[28] A. Garnaev, E. Gluskin, *The widths of a Euclidean ball*, Soviet Math. Dokl. 30 (1984), 200–204.

[29] S. Geman, *A limit theorem for the norm of random matrices*, Ann. Probab. 8 (1980), 252–261.

[30] V. L. Girko, *The circular law*, Theory Probab. Appl. 29 (1984), no. 4, 694–706.

[31] Y. Gordon, *On Dvoretzky's theorem and extensions of Slepian's lemma,* Israel seminar on geometrical aspects of functional analysis (1983/84), II, Tel Aviv Univ., Tel Aviv, 1984.

[32] Y. Gordon, *Some inequalities for Gaussian processes and applications*, Israel J. Math. 50 (1985), 265–289.

[33] Y. Gordon, *Majorization of Gaussian processes and geometric applications*, Probab. Theory Related Fields 91 (1992), 251–267.

[34] F. Götze, A. Tikhomirov, *On the Circular Law*, arXiv:math/0702386.

[35] F. Götze, A. Tikhomirov, *The Circular Law for Random Matrices*, arXiv:0709.3995 , to appear in Ann. Prob.

[36] P. Indyk, *Dimensionality reduction techniques for proximity problems*, Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms, 2000.

[37] K. Johansson, *Shape fluctuations and random matrices*, Comm. Math. Phys. 209 (2000), 437–476.

[38] W. B. Johnson, G. Schechtman, *Very tight embeddings of subspaces of L_p, $1 \leq p < 2$, into l_p^n*, Geom. Funct. Anal. 13 (2003), no. 4, 845–851.

[39] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Ann. Statist. 29 (2001), 295–327.

[40] G. Halász, *Estimates for the concentration function of combinatorial number theory and probability*, Periodica Mathematica Hungarica 8 (1977), 197–211

[41] B. Kashin, *Section of some finite-dimensional sets and classes of smooth functions* (in Russian) Izv. Acad. Nauk. SSSR 41 (1977) 334–351.

[42] A. Kolmogorov, *Sur les propriétés des fonctions de concentrations de M. P. Lévy*, Ann. Inst. H. Poincaré 16 (1958), 27–34.

[43] J. Kahn, J. Komlós, E. Szemerédi, *On the probability that a random* ± 1*-matrix is singular*, J. Amer. Math. Soc. 8 (1995), 223–240.

[44] J. Komlós, *On the determinant of* $(0,1)$ *matrices*, Studia Sci. Math. Hungar. 2 (1967), 7–22.

[45] J. Komlós, *On the determinant of random matrices*, Studia Sci. Math. Hungar. 3 (1968), 387–399.

[46] R. Latala, *Some estimates of norms of random matrices*, Proc. Amer. Math. Soc. 133 (2005), 1273–1282.

[47] M. Ledoux, *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, RI, 2001.

[48] M. Ledoux, *Deviation inequalities on largest eigenvalues*, Geometric aspects of functional analysis, 167–219, Lecture Notes in Math., 1910, Springer, Berlin, 2007.

[49] M. Ledoux, M. Talagrand, *Probability in Banach spaces. Isoperimetry and processes.* Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 23. Springer-Verlag, Berlin, 1991.

[50] P. Lévy, *Théorie de l'addition des variables aléatoires*, Gauthier-Villars, 1937.

[51] W. V. Li, Q.-M. Shao, Gaussian processes: inequalities, small ball probabilities and applications. Stochastic processes: theory and methods, 533–597, Handbook of Statistics, 19, North-Holland, Amsterdam, 2001.

[52] J. E. Littlewood, A. C. Offord, *On the number of real roots of a random algebraic equation. III.* Rec. Math. [Mat. Sbornik] N.S. 12 (54), (1943), 277–286

[53] A. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Adv. Math. 195 (2005), 491–523.

[54] A. E. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann, R. Vershynin, *Euclidean embeddings in spaces of finite volume ratio via random matrices*, J. Reine Angew. Math. 589 (2005), 1–19.

[55] V. A. Marchenko, L. A. Pastur, *The distribution of eigenvalues in certain sets of random matrices*, Mat. Sb., 72 (1967), 507–536.

[56] M. L. Mehta, *Random matrices and the statistical theory of energy levels*. Academic Press, New York-London 1967.

[57] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, *Uniform uncertainty principle for Bernoulli and subgaussian ensembles*, Constr. Approx. 28 (2008), 277–289.

[58] V. Milman, *A new proof of the theorem of A. Dvoretzky on sections of convex bodies*, Funct. Anal. Appl. 5 (1971), 28–37.

[59] V. D. Milman, G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces*. With an appendix by M. Gromov. Lecture Notes in Math. 1200, Springer-Verlag, Berlin 1986.

[60] L. A Pastur, *On the spectrum of random matrices*, Teoret. Mat. Fiz. 10 (1973), 102–112.

[61] V. V. Petrov, *Sums of independent random variables*. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82. Springer-Verlag, New York-Heidelberg, 1975.

[62] M. Rudelson, *Lower estimates for the singular values of random matrices*, C. R. Math. Acad. Sci. Paris 342 (2006), 247–252.

[63] M. Rudelson, *Invertibility of random matrices: norm of the inverse*, Annals of Mathematics 168 (2008), 575–600.

[64] M. Rudelson, R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Communications on Pure and Applied Mathematics 61 (2008), 1025–1045.

[65] M. Rudelson, R. Vershynin, *The Littlewood-Offord problem and invertibility of random matrices*, Adv. Math. 218 (2008), 600–633.

[66] M. Rudelson, R. Vershynin, *The least singular value of a random square matrix is* $O(n^{-1/2})$, Comptes rendus de l'Académie des sciences - Mathématique 346 (2008), 893–896.

[67] M. Rudelson, R. Vershynin, *Smallest singular value of a random rectangular matrix*, Comm. Pure Appl. Math. 62 (2009), 1707–1739.

[68] A. Sankar, D. A. Spielman, S.-H. Teng, *Smoothed analysis of the condition numbers and growth factors of matrices*, SIAM J. Matrix Anal. Appl. 28 (2006), 446–476.

[69] A. Sárközy, E. Szeméredi, *Über ein Problem von Erdös und Moser*, Acta Arithmetica 11 (1965), 205–208

[70] G. Schechtman, *Random embeddings of Euclidean spaces in sequence spaces*, Israel Journal of Mathematics 40, No. 2, 1981, 187–192.

[71] J. Silverstein, *The smallest eigenvalue of a large-dimensional Wishart matrix*, Ann. Probab. 13 (1985), 1364–1368.

[72] D. Slepian, *The one-sided barrier problem for Gaussian noise*, Bell System Tech. J. 41 (1962), 463–501.

[73] S. Smale, *On the efficiency of algorithms of analysis*, Bull. Amer. Math. Soc. (N.S.) 13 (1985), 87–121

[74] A. Soshnikov, *A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices*, J. Statist. Phys. 108 (2002), 1033–1056.

[75] D. Spielman, S.-H. Teng, *Smoothed analysis of algorithms*. Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002), 597–606, Higher Ed. Press, Beijing, 2002

[76] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Hautes Études Sci. Publ. Math. No. 81 (1995), 73–205.

[77] M. Talagrand, *A new look at independence*, Ann. Probab. 24 (1996), 1–34.

[78] T. Tao, V. Vu, *On random* ± 1 *matrices: singularity and determinant*, Random Structures and Algorithms 28 (2006), 1–23.

[79] S. Szarek, *Condition numbers of random matrices*, J. Complexity 7 (1991), 131–149.

[80] T. Tao, V. Vu, *Additive combinatorics.* Cambridge Studies in Advanced Mathematics, 105. Cambridge University Press, Cambridge, 2006.

[81] T. Tao, V. Vu, *On the singularity probability of random Bernoulli matrices*, J. Amer. Math. Soc. 20 (2007), 603–628.

[82] T. Tao, V. Vu, *Random matrices: the circular law*, Commun. Contemp. Math. 10 (2008), no. 2, 261–307.

[83] T. Tao, V. Vu, *Inverse Littlewood-Offord theorems and the condition number of random discrete matrices*, Annals of Math. 169 (2009), 595–632.

[84] T. Tao, V. Vu, *From the Littlewood-Offord problem to the circular law: universality of the spectral distribution of random matrices*, Bull. Amer. Math. Soc. 46 (2009), 377–396.

[85] T. Tao, V. Vu (with appendix by M. Krishnapur), *Random matrices: Universality of ESDs and the circular law*, arXiv:0808.4898, to appear in Ann. Prob.

[86] T. Tao, V. Vu, *A sharp inverse Littlewood-Offord theorem*, Random Structures and Algorithms, to appear

[87] T. Tao, V. Vu, *Random matrices: the distribution of smallest singular values*, Geom. Funct. Anal., to appear

[88] T. Tao, V. Vu, *The Littlewood-Offord problem in high dimensions and a conjecture of Frankl and Füredi*, preprint

[89] C. Tracy, H. Widom, *Level-Spacing Distributions and the Airy Kernel*, Comm. Math. Phys. 159 (1994), 151–174.

[90] C. Tracy, H. Widom, *On orthogonal and symplectic matrix ensembles*, Comm. Math. Phys. 177 (1996), no. 3, 727–754.

[91] C. Tracy, H. Widom, *Distribution functions for largest eigenvalues and their applications*, Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002), 587–596, Higher Ed. Press, Beijing, 2002.

[92] J. von Neumann, H. H. Goldstine, *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc. 53 (1947), 1021–1099.

[93] R. Vershynin, *Spectral norm of products of random and deterministic matrices*, Probability Theory and Related Fields, DOI: 10.1007/s00440-010-0281-z.

[94] R. Vershynin, *How close is the sample covariance matrix to the actual covariance matrix?* preprint.

[95] K. W. Wachter, *Strong limits of random matrix spectra for sample covariance matrices of independent elements*, Annals of Probability 6 (1978), 1–18.

[96] P. Wigner, *On the distribution of the roots of certain symmetric matrices*, Annals of Mathematics 67 (1958), 325–327.

[97] Y. Q. Yin, Z. D. Bai, P. R. Krishnaiah, *On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix*, Probab. Theory Related Fields 78 (1988), 509–521.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Free probability, Planar algebras, Subfactors and Random Matrices

Dimitri Shlyakhtenko*

Abstract

To a planar algebra $\mathcal{P}$ in the sense of Jones we associate a natural non-commutative ring, which can be viewed as the ring of non-commutative polynomials in several indeterminates, invariant under a symmetry encoded by $\mathcal{P}$. We show that this ring carries a natural structure of a non-commutative probability space. Non-commutative laws on this space turn out to describe random matrix ensembles possessing special symmetries. As application, we give a canonical construction of a subfactor and its symmetric enveloping algebra associated to a given planar algebra $\mathcal{P}$. This talk is based on joint work with A. Guionnet and V. Jones.

Mathematics Subject Classification (2010). Primary: 46L37, 46L54; Secondary 15B52.

Keywords. Free probability, von Neumann algebra, random matrix, subfactor, planar algebra.

1. Introduction

The aim of this paper is to explore the appearance of planar algebra structure in three areas of mathematics: subfactor theory; free probability theory; and random matrices.

Jones' subfactor theory has lead to a revolution in understanding what may be termed "quantum symmetry". The standard invariant of a subfactor — the so-called lattice of higher relative commutants, or "λ-lattice" [Pop95, GHJ89] is a remarkable mathematical object, which can represent a very general type of symmetry. For example, a subfactor inclusion (and so its standard invariant)

*Research supported by NSF grants DMS0555680 and DMS0900776.
Department of Mathematics, UCLA, Los Angeles, CA 90095, USA.
E-mail: shlyakht@math.ucla.edu.

can be associated to a Lie group representation. In this case, the vector spaces that make up the standard invariant are the spaces of intertwiners between tensor powers of that representation. Thus the standard invariant of such a subfactor can be used to encode the representation theory of a Lie group, and thus symmetries associated with Lie group actions.

In his groundbreaking paper [Jon99, Jon01] Jones (building on an earlier algebraic axiomatization of standard invariants by Popa [Pop95]) showed that there is a striking way to characterize standard invariants of subfactors: these are exactly *planar algebras* (see §3.4 below for a definition). Very roughly, one can think of a planar algebra as a sequence of vector spaces consisting of vectors invariant under some "quantum symmetry", together with very general ways (dictated by planar diagrams) of producing new invariant vectors from existing vectors. The planar algebra thus *encodes* the underlying symmetry. In the context of the present paper, we shall use the terms "quantum symmetry" and "planar algebra" interchangeably.

Curiously, planar diagrams also occur in random matrix theory. Certain random multi-matrix ensembles (see §4.7 below) are asymptotically described by combinatorics involving counting *planar maps* (these objects are very much like planar diagrams appearing in the definition of planar algebras). This fact has been discovered and extensively used by physicists, starting from the works of 't Hooft, Brezin, Iszykson, Parisi, Zuber and others (see e.g. [tH74, BIPZ78]). A rigorous proof of convergence was obtained by Guionnet and Maurel Segala (see [Gui06, GMS06] and references therein) and Ercolani and McLaughlin [EM03].

Finally, turning to Voiculescu's free probability theory [VDN92], it was shown by Speicher [Spe94] and others that many important free probability laws (such as the semicircle law, the free Poisson law and so on) have combinatorial descriptions involving counting planar objects (such as non-crossing partitions, which are also very closely related to planar diagrams).

Thus one is faced with two natural questions. First, why do these planar structures appear in these three areas? And second, how can these similarities be exploited?

Concerning the first question, we do not know a fully satisfactory answer. However, if one grants that planar structure is necessary to describe "quantum symmetries" (i.e., subfactors), then one is able to find explanations for appearances of planar structure in free probability theory and random matrices. We show that one has a natural notion of a *non-commutative probability law having a quantum symmetry* — this law is given by a trace on a ring naturally associated to a planar algebra. Mathematically, this is accomplished by a "change of rings" procedure, where we replace the ring of non-commutative polynomials in K variables with a certain canonical ring associated to a given planar algebra (see §3.9). This "change of rings" is analogous to the passage from some probability space Ω to the quotient space Ω/G in the case that the

laws of some family of random variables are invariant under the action of a group G.

Also, we show how to construct random matrix ensembles, which asymptotically give rise to a non-commutative law with a given quantum symmetry.

This means that any time one considers a natural equation in free probability theory, or a natural equation giving the asymptotics of a random matrix ensemble, this equation must make sense not only as an equation involving polynomials in K non-commuting indeterminates, but also arbitrary planar algebra elements. Thus the equation (and so its solutions) must have a natural planar structure.

Concerning the second question, we give a number of applications of our techniques. One such application is a version of the ground-breaking theorem of Popa [Pop95, PS03] which states that every planar algebra $\mathcal{P}$ arises from a subfactor $N \subset M$ with N, M isomorphic to free group factors. It turns out that both N and M can in fact be chosen to be natural non-commutative probability spaces "in the presence of the symmetry $\mathcal{P}$". On the random matrix side, our approach gives a mathematical framework to formulate the work of a number of physics authors [EZJ92, Kos89, ZJ03] on the so-called $O(n)$ matrix model. In fact, using our techniques one can make rigorous sense of the $O(n)$ matrix model for $n \in \{2\cos\frac{\pi}{n} : n \geq 3\} \cup [2, +\infty)$ (non-integer values of n are used in the physics literature).

The remainder of the paper is organized as follows. We first discuss some basic notions from free probability theory and subfactors. Next, we discuss a notion of a non-commutative probability law having a symmetry encoded by a planar algebra $\mathcal{P}$ and present some applications to subfactor theory. Finally, we show that one can construct random matrix ensembles that model certain non-commutative laws with a given planar algebra symmetry $\mathcal{P}$, and explain connections with a class of random matrix ensembles used in the physics literature, and derive some random matrix consequences.

This paper is based on the joint work with A. Guionnet and V.F.R. Jones [GJS08, GJS09].

2. Background and Basic Notions: Free Probability and Non-commutative Probability Spaces

2.1. Non-commutative probability spaces. Recall (see for example [VDN92]) that an algebraic non-commutative probability space $(A, 1_A, \tau)$ consists of an algebra A with unit 1_A and a unital linear functional $\tau : A \to \mathbb{C}$. We often make the assumption that A is a $*$-algebra and τ is a *trace*, i.e., $\tau(ab) = \tau(ba)$ for all $a, b \in A$. Elements of A are called *non-commutative random variables*. Here are a few examples:

Example 2.2. (a) If $(\mathfrak{X}, \mu)$ is a measure space and μ is a probability measure, then $(A = L^\infty(\mathfrak{X}, \mu), 1_A, f \xrightarrow{\tau} \int f d\mu)$ is a non-commutative probability space.

(b) For any N, the algebra of $N \times N$ matrices $(A = M_{N\times N}(\mathbb{C}), 1_A = \mathrm{Id}, \tau = \frac{1}{N}Tr)$ is a non-commutative probability space.

(c) Consider $A = M_{N\times N}(L^{\infty,-}(\mathfrak{X}, \mu))$, with $(\mathfrak{X}, \mu)$ as in (a). Thus elements of A are *random matrices*. Then $(A, 1_A, \mathbb{E}(\frac{1}{N}Tr(\cdot)))$ is a non-commutative probability space.

Note that in all of these examples, τ is a trace: $\tau(xy) = \tau(yx)$.

In order to be able to do analysis on non-commutative probability spaces we make the assumption that the algebra $(A, 1_A, \tau)$ is represented (by bounded or unbounded operators) on a Hilbert space H by a faithful unital representation π, so that $\tau(a) = \langle \Omega, \pi(a)\Omega \rangle$ for some fixed vector $\Omega \in H$.

Elements of non-commutative probability spaces are called non-commutative random variables.

2.3. Non-commutative laws. Given $K = 1, 2, \ldots$ classical real random variables $X_1, \ldots, X_K$, which we can think of as an $\mathbb{R}^K$-valued function X on some probability space $(\mathfrak{X}, \mu)$, their joint law is defined to be the push-forward by $\tau = X_*\mu$ of μ to a probability measure on $\mathbb{R}^K$. If μ has finite moments, we obtain a linear functional on the algebra of polynomials on $\mathbb{R}^K$.

By analogy, given non-commutative random variables $X_1, \ldots, X_K \in A$, their *non-commutative law* $\tau_{X_1,\ldots,X_K}$ is the linear function on the algebra of all non-commutative polynomials in K indeterminates $\mathbb{C}[t_1, \ldots, t_K]$ obtained by composing τ with the canonical map sending t_j to X_j. In other words

$$\tau_{X_1,\ldots,X_K}(P(t_1, \ldots, t_K)) = \tau(P(X_1, \ldots, X_K))$$

for any non-commutative polynomial P.

If $K = 1$, non-commutative laws are the same as commutative laws, modulo identification of measures with linear functionals they induce on polynomials by integration. For example, in the case of a single self-adjoint matrix $Y \in (M_{N\times N}, \frac{1}{N}Tr)$, its non-commutative law corresponds to integration against the measure $\mu_Y = \frac{1}{N}\sum \delta_{\lambda_j}$, where $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of Y. If Y is a random matrix, its non-commutative law captures the expected value of the random spectral measures associated to Y, $\mathbb{E}(\mu_Y)$.

The classical notion of independence of random variables can be reformulated algebraically by stating that $(X_1, \ldots, X_K)$ is independent from $(X_{K+1}, \ldots, X_{K+L})$ in a non-commutative probability space (A, τ) if the law of $(X_1, \ldots, X_{K+L}) \in (A, \tau)$ is the same as that of the variables

$$(\alpha_1(X_1), \ldots, \alpha_1(X_K), \alpha_2(X_{K+1}), \ldots, \alpha_2(X_{K+L})) \in (A \otimes A, \tau \otimes \tau).$$

Here $\alpha_1(X) = X \otimes 1$, $\alpha_2(X) = 1 \otimes X$ are two natural embeddings of A into $A \otimes A$.

Voiculescu developed his *free probability theory* (see e.g. [VDN92]) around another notion of independence, free independence. For this notion, we say that $(X_1, \ldots, X_K)$ is freely independent from $(X_{K+1}, \ldots, X_{K+L})$ in a non-commutative probability space (A, τ) if the law of $(X_1, \ldots, X_{K+L}) \in (A, \tau)$ is the same as that of the variables

$$(\alpha_1(X_1), \ldots, \alpha_1(X_K), \alpha_2(X_{K+1}), \ldots, \alpha_2(X_{K+L})) \in (A * A, \tau * \tau),$$

where $*$ denotes the free product [Voi85, VDN92], and α_1, α_2 are the natural embeddings of A into $A * A$ (into the first and second copy, respectively).

If τ is a non-commutative law satisfying positivity and boundedness requirements, the GNS construction yields a representation of $\mathbb{C}[t_1, \ldots, t_K]$ on $L^2(\tau)$ and thus generates a von Neumann algebra $W^*(\tau)$. The non-commutative case here differs significantly from the commutative case. In the commutative case, $W^*(\tau) = L^\infty(\mathfrak{X})$, and, notably, all measure spaces $\mathfrak{X}$ are isomorphic (at least for laws τ which are non-atomic). In the non-commutative case, the von Neumann algebras $W^*(\tau)$ are much more diverse, and it is in general a very difficult and challenging question to decide, for two laws τ, τ', when $W^*(\tau) \cong W^*(\tau')$, or to somehow identify the isomorphism class of $W^*(\tau)$.

3. Symmetries: Subfactors, Planar Algebras, and Non-commutative Laws

3.1. Non-commutative laws with quantum symmetry. Consider a complex-valued classical random variable Z; thus we actually have a pair of random variables $Z, \bar{Z}$, whose joint law is described by a probability measure μ on $\mathbb{C} = \mathbb{R}^2$: for any function of two variables $f(x, y)$, we are interested in the value

$$\iint f(z, \bar{z}) d\mu(z, \bar{z}).$$

In this way, the law of $(Z, \bar{Z})$ is a functional on the space of functions on $(-\infty, \infty) \times (-\infty, \infty)$.

Assume that we know that the law of $(Z, \bar{Z})$ is invariant under rotations: $(Z, \bar{Z}) \sim (wZ, \bar{w}\bar{Z})$ for any $w \in \mathbb{C}$, $|w| = 1$. Then the joint law of $(Z, \bar{Z})$ is completely determined by its "radial part", the integrals of the form

$$\int g(|z|) d\mu(z, \bar{z}),$$

and thus defines a linear functional on the space of rotation-invariant functions, i.e., effectively on the space of functions on $[0, +\infty) = \mathbb{C}^2/\text{rotation}$.

Thus the *presence of a symmetry dictates that we use a different probability space*. Our aim is to extend this observation to the non-commutative setting, allowing the most general notions of symmetry possible.

We defined a non-commutative probability law to be a linear functional τ defined on the algebra $A = \mathbb{C}[X_1, \ldots, X_K]$ of non-commutative polynomials in K variables. If symmetries are present, this choice of the algebra A may not be suitable. In this case the algebra A (the non-commutative analog of the ring of polynomials on $\mathbb{R}^K$) must be replaced by the analog of the ring of functions on a different algebraic variety. For instance, one may be interested in $*$-probability spaces, i.e., we want to have an algebra A that has a non-trivial adjoint operation (involution). This can be accomplished by considering the algebra $B = \mathbb{C}[X_1, \ldots, X_K, X_1^*, \ldots, X_K^*]$ and defining X_j^* to be the adjoint of X_j. An even more interesting situation is the case that our algebra B has a natural symmetry. For example, we may consider the action of the unitary group $U(K)$ on B given on the generators by

$$U \cdot X_k = \sum U_{ik} X_i, \qquad U \cdot X_k^* = \sum \overline{U_{ik}} X_i^*, \qquad U = (U_{ij}). \tag{3.1.1}$$

In this case we may only be interested in a part of B, the algebra $B^{U(K)}$ consisting of $U(K)$-invariant elements. One can easily see that $B^{U(K)}$ is not even a finitely-generated algebra, but it is the natural non-commutative probability space on which to define $U(K)$-invariant laws.

More generally, in this paper we will be interested in non-commutative laws defined on a class of "symmetry algebras", which are the analogs of algebras such as $B^{U(K)}$ above for more general symmetries (including actions of quantum groups).

As is well-known, subfactor theory of Jones provides a framework for considering such very general symmetries. To formalize our notion of a "non-commutative probability law with a quantum symmetry", we shall first review Jones' notion of planar algebras [Jon99, Jon01].

3.2. The standard invariant of a subfactor: spaces of intertwiners. Planar algebras [Jon99, Jon01] were introduced by Jones in his study of invariants of subfactors of II_1 factors.

Let $M_0 \subset M_1$ be an inclusion of II_1 factors of finite Jones' index [Jon83, GHJ89]. Then M_1 can be regarded as a bimodule over M_0 by using the left and right multiplication action of M_0 on M_1. Using the operation of the relative tensor product of bimodules (see e.g. [Con, Pop86, Con94, Bis97]) one can construct other M_0, M_0-bimodules by considering tensor powers

$$M_k = \underbrace{M_1 \otimes_{M_0} \otimes \cdots \otimes_{M_0} M_1}_{k}.$$

One can then consider the intertwiner spaces

$$A_{ij} = \mathrm{Hom}_{M_0, M_0}(M_i, M_j)$$

consisting of all homomorphisms from M_i to M_j, which are linear for both the left and the right action of M_0. Because the index of $M_0 \subset M_1$ is finite, these spaces turn out to be finite-dimensional. The system of intertwiner spaces A_{ij} has more structure than the algebra structure of the individual A_{ij}'s. For example, having an intertwiner $T : M_i \to M_j$ one can also construct an "induced representation" intertwiner $T \otimes 1 : M_{i+1} \to M_{j+1}$. More generally, one can restrict intertwiners, take their tensor products, etc., thus providing many operations involving elements of the various A_{ij}'s.

The following example explains how classical representation theory of a Lie group can be viewed in subfactor terms. Similar examples exist also in the case of quantum group representations:

Example 3.3. Let G be a Lie group and V be an irreducible representation of G, and denote by V^{op} the representation on the dual of V. Let M be a II_1 factor carrying an action of G satisfying a technical condition of being properly outer (such an action always exists with M a hyperfinite II_1 factor or a free group factor). Consider the "Wassermann-type" inclusion

$$M_0 = M^G \subset (M \otimes End(V))^G = M_1.$$

Here N^G denotes the fixed points algebra for an action of G on N, and G acts on $End(V)$ by conjugation. Then

$$\mathrm{Hom}_{M_0,M_0}(M_k) = \mathrm{Hom}_G(\underbrace{V \otimes V^{op} \otimes \cdots \otimes V \otimes V^{op}}_{k})$$

is the space of all G-invariant linear maps on $(V \otimes V^{op})^{\otimes k}$.

The main theorem of Jones [Jon99, Jon01] is that there is a beautiful abstract characterization of systems of intertwiner spaces associated to a subfactor (also called "standard invariants", "λ-lattices", systems of higher-relative commutants): such systems are exactly the *planar algebras*. His proof relied on an earlier axiomatization of λ-lattices by Popa [Pop95].

3.4. Planar algebras.

To state the definition of a planar algebra, let us introduce the notion of a *planar tangle* T with r input disks or sizes $k_1, \ldots, k_r$ and output disk of size k (we'll write $\mathcal{T}(k_1, \ldots, k_r; k)$ for the set of such tangles). Such a tangle is given by drawing (up to isotopy on the plane) r "input" disks $(D_j : j = 1, \ldots, r)$ inside the "output" disk D. Each disk D_l has $2k_l$ points marked on its boundary (one of which is marked as the "first" point). The output disk D has $2k$ points marked on its boundary, one of which is marked "first". Furthermore, all marked boundary points are connected to other marked points by non-crossing paths.[1]

[1] One also assumes that the connected components of $D \setminus \bigcup_j D_j$ are colored by two colors, so that adjacent regions are colored by different colors. We shall, however, ignore this part of this structure in this paper.

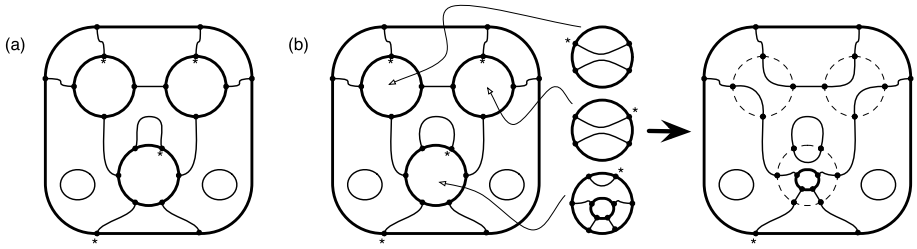


Figure 1. Planar tangles; composing planar tangles.

Figure 1(a) shows an example of a planar tangle in $\mathcal{T}(3,3,2;3)$; the first point on each interior disk is labeled by a $*$. Note that tangles may contain loops which are not connected to any interior disks.

Tangles can be composed by gluing the output disk of one tangle inside an input disk of another tangle in a way that aligns points marked "first" and preserves the orientation of boundaries (see Figure 1(b), which illustrates the composition of a tangle in $\mathcal{T}(3,2,2;3)$ with three tangles, from $\mathcal{T}(2;3)$, $\mathcal{T}(;2)$ and $\mathcal{T}(;2)$). (This is only possible if disks are of matching sizes).

Definition 3.5. Let $(P_k : k = 0,1,2,\dots)$ be a collection of vector spaces. We say that $(P_k)_{k\geq 0}$ forms a planar algebra if any planar tangle $T \in \mathcal{T}(k_1,\dots,k_r;k)$ gives rise to a multi-linear operation $Op(T) : P_{k_1} \otimes \cdots \otimes P_{k_r} \to P_k$ in such a way that the assignment $T \to Op(T)$ is natural with respect to composition of tangles and of multilinear maps.

Very roughly, one should think of the spaces P_k as the space of "intertwiners" of degree $2k$ for some quantum symmetry (see §3.6.1 below). The various operations $Op(T)$ correspond to the various ways of combining such intertwiners to form new intertwiners.

We also often make the assumption that the space P_0 is one-dimensional and all P_k are finite-dimensional. In particular, a tangle T with no input disks and one output disk with zero marked points and no paths inside gives rise to a basis element of P_0, which we'll denote by $\emptyset$. If we instead consider a tangle T' with no input disks, one output disk with no marked points, and a simple closed loop inside of the output disk, then T' produces an element $\delta\emptyset$ in P_0 (where δ is some fixed number). Furthermore, it follows from naturality of composition of tangles that if some tangle T is obtained from a tangle T' by removing a closed loop, then $Op(T) = \delta Op(T')$.

The tangle in Figure 2(a) gives rise to a bilinear form on each A_k, which we assume to be non-negative definite. We endow each P_k with an involution compatible with the action of orientation-preserving planar maps on tangles. Finally, we assume a spherical symmetry, so that we consider tangles up to isotopy on the sphere (and not just the plane).

A planar algebra satisfying these additional requirements is called a *subfactor planar algebra* with parameter δ. It is a famous result of Jones [Jon83] that $\delta \in \{2\cos\frac{\pi}{n} : n \geq 3\} \cup [2,+\infty)$, and all of these values can occur.

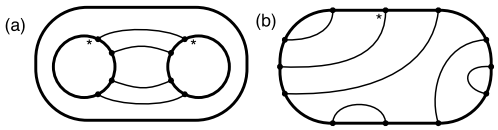


Figure 2. Canonical bilinear form; Temperley Lieb diagrams.

3.6. Examples of planar algebras. Planar algebras can be thought of as families of linear spaces consisting of vectors "obeying a symmetry", where the word symmetry is taken in a very generalized sense (such "symmetries" include group actions as well as quantum group actions). We consider a few examples:

3.6.1. Planar algebras of polynomials. Let $X_1, \ldots, X_K, X_1^*, \ldots, X_K^*$ be indeterminates, and denote by A the algebra spanned by alternating monomials of the form $X_{i_1} X_{j_1}^* \cdots X_{i_k} X_{j_k}^*$. Let P_k be the linear subspace of A consisting of all elements that have degree $2k$. We claim that $\mathcal{P} = (P_k)_{k \geq 0}$ is a planar algebra if endowed with the following structure. Given a monomial $W = X_{i_1} X_{j_1}^* \cdots X_{i_k} X_{j_k}^* \in P_k$, associate to it the labeled disk $D(W)$ whose $2k$ boundary points are labeled (clockwise, from the "first" point) by the $2k$-tuple $(i_1, j_1, i_2, j_2, \ldots, i_k, j_k)$. Now given a planar tangle $T \in \mathcal{T}(k_1, \ldots, k_r; k)$ and monomials $W_1, \ldots, W_r$ of appropriate degrees, we define

$$Op(T)(W_1, \ldots, W_r) = \sum_W C_W W.$$

Here the sum is over all monomials $W \in A_k$ and C_W are integers obtained as follows. Glue the disks $D(W_j)$ into the input disks of T and then the output disk of T into $D(W)$. We obtain a collection of disks, whose marked boundary points are connected by curves. Then C_W is the total number of ways to assign integers from $\{1, \ldots, K\}$ to these curves, so that each curve has the same label as its endpoints. ($C_W = 0$ if no such assignment exists).

In this case, $\mathcal{P}$ is actually a subfactor planar algebra with parameter $\delta = K$ (the number of ways to assign an integer from $\{1, \ldots, K\}$ to a closed loop). The corresponding subfactor inclusion is rather trivial: it corresponds to the $K \times K$ matrix inclusion $M_0 = M \subset M \otimes M_{K \times K}(\mathbb{C}) = M_1$, for any II_1 factor M.

Consider the action of the unitary group $U(K)$ on each P_k defined by (3.1.1). In other words, we identify P_k with the k-th tensor power of $\mathbb{C}^K \otimes \overline{\mathbb{C}^K} = \mathrm{End}(\mathbb{C}^K)$, where $\mathbb{C}^K$ is the basic representation of $U(K)$. Then the linear subspaces $P_k^{U(K)}$ consisting of vectors fixed by the $U(K)$ action turn out to form a planar algebra $\mathcal{P}^{U(K)}$ (taken with the restriction of the planar algebra structure of $\mathcal{P}$). The associated subfactor has the form

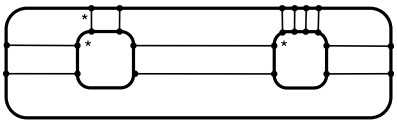
$$M^{U(K)} \subset (M \otimes \mathrm{End}(\mathbb{C}^K))^{U(K)}.$$

3.6.2. The Temperley-Lieb planar algebra. Let TL_k be the linear space spanned by tangles $T \in \mathcal{T}(;k)$ with *no* internal disks and $2k$ points on the outer disk. Such tangles are called *Temperley-Lieb diagrams* (see Figure 2(b)). Then $TL = (TL_k)_{k\geq 0}$ is a planar algebra in the following natural way. Given any tangle $T \in \mathcal{T}(k_1, \ldots, k_r; k)$ and Temperley-Lieb diagrams $T_1, \ldots, T_r$, $Op(T)(T_1, \ldots, T_r)$ is defined to be the result of gluing the diagrams $T_1, \ldots, T_r$ into the input disks of T, provided that we agree that closed loops contribute a multiplicative factor of δ. TL is actually a subfactor planar algebra when δ is in the set of allowed index values $\{2\cos\frac{\pi}{n} : n \geq 3\} \cup [2, +\infty)$.

It should be noted that *any* planar algebra $\mathcal{P}$ contains a homomorphic image of TL; indeed, TL elements arise as $Op(T)$ when $T \in \mathcal{T}(;k)$.

3.7. Algebras and non-commutative probability spaces arising from planar algebras.

A planar algebra $\mathcal{P} = (P_k)_{k\geq 0}$ has, by definition, a large variety of mutli-linear operations. We shall single out the following bilinear operations $\wedge_k$, each of which is an associative multiplication on $\oplus_{n\geq k} P_k$. The operation $\wedge_k$ takes $P_{k+n} \times P_{k+m} \to P_{k+m+n}$ and is given by the following tangle (here $k = 2$, $n = 1$ and $m = 2$):



3.7.1. The product $\wedge_0$. Perhaps the easiest way to see the importance of these operations is to realize that in the case of planar algebra of polynomials (see §3.6.1) the multiplication $\wedge_0$ is just the ordinary multiplication of polynomials.

Thus if we think of $\oplus_{k\geq 0} P_k$ as a linear space consisting of vectors which are invariant under some "quantum symmetry", the product $\wedge_0$ is a kind of tensor product of these invariants, and thus $(\mathcal{P}, \wedge_0)$ has the natural interpretation of the algebra of "invariant polynomials".

3.7.2. The higher products $\wedge_k$. In the case of the polynomial algebra (§3.6.1), the product $\wedge_k$ corresponds to the product on the algebra of differential operators of degree k. Let us consider such operators of the form (for simplicity, if k is even)

$$X_{i_1} X^*_{j_1} \cdots X_{i_{k/2}} X^*_{j_{k/2}} X_{t_1} X^*_{s_1} \cdots X_{t_n} X^*_{s_n} \partial_{X_{i_{k/2+1}}} \partial_{X^*_{j_{k/2+1}}} \cdots \partial_{X_{i_k}} \partial_{X^*_{j_k}} \in P_{k+n}.$$

Such expressions can be multiplied using the convention that $\partial_{X^a_s} X^b_t = \delta_{a\neq b}\delta_{s=t} 1$, where $a, b \in \{\ , *\}$. This is exactly the multiplication $\wedge_k$.

Note that the map $\mathcal{E}_k$ given by the tangle in Figure (3)(c) defines a natural map from $(\mathcal{P}, \wedge_k)$ to $(\mathcal{P}, \wedge_0)$.

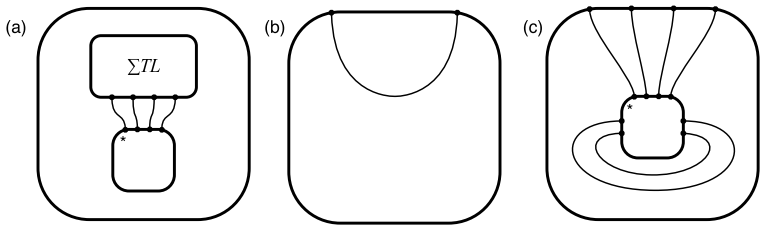


Figure 3. (a) The Voiculescu trace; here $\sum TL$ stands for the sum of all TL elements with the appropriate number of strings. (b) The element $\cup$. (c) The map $\mathcal{E}_k$ (here $k = 2$).

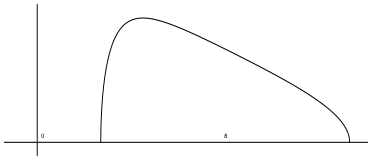


Figure 4. Free Poisson law ($\delta = 8$).

Definition 3.8. A *planar algebra law* associated to a planar algebra $\mathcal{P}$ is a linear functional τ on the algebra $(\mathcal{P}, \wedge_0)$, *so that* $\tau \circ \mathcal{E}_k$ *is a trace on* $(\mathcal{P}, \wedge_k)$ *for any* $k \geq 0$.

Since P_k can be thought of as the space of vectors with a "quantum symmetry encoded by $\mathcal{P}$", a planar algebra law is a law having this "quantum symmetry".

3.9. The Voiculescu trace on $(\mathcal{P}, \wedge_0)$.

Any planar algebra probability space comes with a natural trace $\tau = \tau_{TL}$ given by the tangle in Figure (3)(a).

Lemma 3.10. *[GJS08] (Non-commutative analog of the* χ*-squared distribution). Consider the element* $\cup \in TL$ *described in Figure* (3)*(b). Then law of* $\cup \in TL \subset (\mathcal{P}, \wedge_0, \tau_{TL})$ *is the free Poisson law of parameter* δ *(see Figure 4).*

The polynomial planar algebra (see §3.6.1) contains TL; one can compute that $\cup = \sum_{i=1}^{K} X_i X_i^*$, which explains the analogy with the χ-squared law.

Theorem 3.11. *[GJS08] Assume that* $\mathcal{P}$ *is a subfactor planar algebra. Then trace* τ_{TL} *is non-negative definite. If* $\delta > 1$*, then the von Neumann algebra* $M_0(\mathcal{P}) = W^*(\tau_{TL})$ *generated in the GNS representation is a* II_1 *factor.*

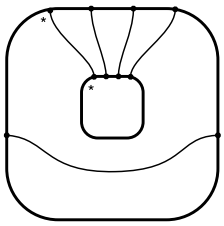
There are several ways in which one can obtain this statement. One such way is show explicitly that the Hilbert space $L^2(\tau_{TL})$ can be identified with the L^2 direct sum of the spaces making up the planar algebra [JSW08]. To prove that $M_0(\mathcal{P})$ is a factor, one first shows that the element $\cup$ generates a maximal

abelian sub-algebra. Thus the center of M is contained in $W^*(\cup)$; some further analysis shows that the center is in fact trivial.

In a similar way one can prove:

Theorem 3.12. *[GJS08] For a subfactor planar algebra $\mathcal{P}$*, consider *the trace τ_{TL}^n on $(\mathcal{P}, \wedge_n)$ given by $\tau_{TL} \circ \mathcal{E}_n$. Then τ_{TL}^n is non-negative definite, and the von Neumann algebra $M_n(\mathcal{P}) = W^*(\tau_{TL}^n)$ is a II_1 factor whenever $\delta > 1$.*

3.13. Application: constructing a subfactor realizing a given planar algebra. The following tangle gives rise to a natural inclusion from $M_0(\mathcal{P})$ into $M_1(\mathcal{P})$:



It turns out that this makes $M_0(\mathcal{P})$ into a finite-index subfactor of $M_1(\mathcal{P})$, which *canonically realizes $\mathcal{P}$:*

Theorem 3.14. *[GJS08] (a) The inclusions $M_0(\mathcal{P}) \subset M_1(\mathcal{P}) \subset \cdots \subset M_{n-1}(\mathcal{P}) \subset M_n(\mathcal{P})$ are canonically isomorphic to the tower of basic constructions for $M_0(\mathcal{P}) \subset M_1(\mathcal{P})$. (b) The planar algebra associated to the inclusion $M_0(\mathcal{P}) \subset M_1(\mathcal{P})$ is again $\mathcal{P}$.*

In other words, we are able to construct a canonical subfactor realizing the given planar algebra. A construction that does this was given earlier by Popa [Pop93, Pop95, Pop02, PS03] using amalgamated free products. In fact, it turns out that our construction is related to his; in particular, the algebras $M_i(\mathcal{P})$ are isomorphic to certain amalgamated free products [GJS09, KS09a, KS09b]. We are able to identify the isomorphism classes of the algebras $M_j(\mathcal{P})$:

Theorem 3.15. *[GJS09, KS09a, KS09b] Assume that* $\dim P_0 = \mathbb{C}$, *$\delta > 1$ and $\mathcal{P}$ is finite-depth of global index I. Then*

$$M_0(\mathcal{P}) \cong L(\mathbb{F}_t)$$

where $t = 1 + 2(\delta - 1)I$. More generally, $M_j(\mathcal{P}) = L(\mathbb{F}_{t_j})$ with $t_j = 1 + \delta^{-2j}(\delta - 1)I$, $j \geq 0$.

Here $L(\mathbb{F}_t)$ is the interpolated free group factor [Dyk94, Răd94]: $L(\mathbb{F}_t) = pL(\mathbb{F}_n)p$ where p is a projection so that $t - 1 = \tau(p)^2(n-1)$.

Of course, it should be noted that rather than considering von Neumann algebras $M_j(\mathcal{P}) = W^*(\mathcal{P}, \wedge_j, \tau_{TL} \circ E_j)$ one can also consider the C^*-algebras $C^*(\mathcal{P}, \wedge_j, \tau_{TL} \circ E_j)$. Little is known about their structure.

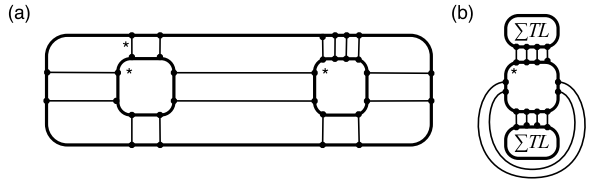


Figure 5. (a) The multiplication $\boxtimes_k$ (there are k horizontal lines joining the input disks). (b) The trace $\tau \boxtimes_k \tau$ (there are k loops).

3.16. Application: the symmetric enveloping algebra. Consider the associative multiplication $\boxtimes_k$ defined on $\oplus_{n\geq k} P_k$ by the tangle in Figure 5(a) and the trace $\tau \boxtimes_k \tau$ on $(\bigoplus_{n\geq k} P_k, \boxtimes_k)$ defined in Figure 5(b).

Let us call $M_k \boxtimes M_k$ the von Neumann algebra generated by this algebra in the GNS representation. These algebras are related to Popa's symmetric enveloping algebra $M_1 \boxtimes_{e_0} M_1^{op}$. For $k = 1$ we obtain exactly the symmetric enveloping algebra, at least in the Temperley-Lieb case.

The symmetric enveloping algebra was introduced by Popa as an important analytical tool in the study of the "quantum symmetry" behind a planar algebra. For example, such analytic properties as amenability, property (T) and so on are encoded by the symmetric enveloping algebra [Pop99].

4. Random Matrices and Planar Algebras

4.1. GUE and the Voiculescu trace τ_{TL}. Let $M_{N\times N'}$ denote the linear space of complex $N \times N'$ matrices. Let $K = 1, 2, \ldots$ be an integer, and endow $(M_{N\times N'})^K$ with the Gaussian measure

$$d\mu^{(N,N')}(A_1, \ldots, A_K, A_1^*, \ldots, A_K^*) = \frac{1}{Z_N} \exp(-\frac{1}{2} N Tr(\sum A_j^* A_j))\ dA_1 \cdots dA_K dA_1^* \cdots dA_K^*.$$

Here $dA_j dA_j^*$ stands for Lebesgue measure on the j-th copy of $M_{N\times N'}$.

A K-tuple of matrices $(A_1, \ldots, A_K)$ chosen at random from $(M_{N\times N'})^K$ according to this measure is called the Gaussian Unitary Ensemble (GUE).

Let Q be a non-commutative polynomial in $X_1, \ldots, X_K, X_1^*, \ldots, X_K^*$ which is a linear combination of monomials of the form $X_{i_1} X_{j_1}^* \cdots X_{i_p} X_{j_p}^*$ (in other words, we can think of Q as an element of $(\mathcal{P}, \wedge_0)$, where $\mathcal{P}$ is the planar algebra of polynomials, see §3.6.1). For each N, N', consider the non-commutative law $\tau^{(N,N')}$ defined by

$$\tau^{(N,N')}(Q) = \int \frac{1}{N} Tr(Q(A_1, \ldots, A_K, A_1^*, \ldots, A_K^*)) d\mu^{(N,N')}(A_1, \ldots, A_K, A_1^*, \ldots, A_K^*).$$

The non-commutative law $\tau^{(N,N')}$ captures certain aspects of the random multi-matrix ensemble $(A_1, \ldots, A_K)$. For example, the value of $\tau^{(N)}((A_1A_1^*)^p)$ is the p-th moment of the empirical spectral measure associated to $A_1A_1^*$: if $\lambda_1 < \cdots < \lambda_N$ are the random eigenvalues of $A_1A_1^*$, then

$$\tau^{(N)}((A_1A_1^*)^p) = \mathbb{E}\left(\sum \lambda_j^p\right).$$

In his seminal paper [Voi91], Voiculescu showed that the laws $\tau^{(N)}$ have a limit as $N \to \infty$; rephrasing slightly he proved:

Theorem 4.2. *[Voiculescu] With the above notation, assume that $N, N' \to \infty$ so that $N'/N \to 1$. Then $\tau^{(N)} \to \tau_{TL}$, where τ_{TL} is the Voiculescu trace on the planar algebra of polynomials.*

One can re-derive some well-known random matrix results from this theorem. For example, combining it with Lemma 3.10, one can recover convergence of singular values of block random GUE matrices to the Marcenko-Pastur law [MP67].

4.3. The case of a general planar algebra. It turns out that Theorem 4.2 also holds in the context of more general planar algebras (i.e., "in the presence of symmetry"). We now describe the appropriate random matrix ensembles.

4.3.1. Graph planar algebras. Our construction relies on the following fact [Jon01, GJS08]:

Proposition 4.4. *Every planar algebra $\mathcal{P}$ is a subalgebra (in the sense of planar algebras) of some graph planar algebra $\mathcal{P}^\Gamma$.*

Here the graph planar algebra $\mathcal{P}^\Gamma$ is a planar algebra canonically associated to an arbitrary bipartite graph, taken with its Perron-Frobenius eigenvector μ (if $\mathcal{P}$ is finite depth, Γ can be taken to be a finite graph). The spaces $\mathcal{P}_k^\Gamma$ have as linear bases the sets of closed paths of length $2k$ on Γ. The planar algebra structure is defined in a manner analogous to the case of the polynomial planar algebra, §3.6.1; see [Jon01] for details. The graph Γ can be chosen to be finite if the planar algebra is finite depth (in particular, if $\delta < 2$).

4.4.1. Random matrix ensembles on graphs. Let $\mathcal{P}$ be a planar algebra of finite depth. Thus $\mathcal{P} \subset \mathcal{P}^\Gamma$ for some finite bi-partite graph. Let us write $\mu(v)$ for the value of the Perron-Frobenius eigenvector at a vertex v of Γ.

To an oriented edge e of Γ which starts at v and ends at w we associated a matrix X_e of size $[N\mu(v)] \times [N\mu(w)]$ (here $[\cdot]$ denotes the integer part of a number). To a path $e_1 \cdots e_n$ in the graph we associate the product of matrices $X_{e_1} \cdots X_{e_n}$ (here $X_{e^o} = X_e^*$ if e^o is the edge e but with opposite orientation).

Thus any element $W \in \bigoplus_k P_k$ is a specific expression in terms of the matrices $\{X_e\}_{e\in\mathcal{E}(\Gamma)}$. For example, let $\cup$ be as in Figure 3(b). Then $\cup = \sum_e \sqrt{\frac{\mu(v)}{\mu(w)}} X_e X_e^*$, the sum taken over all positively oriented edges; here v and w are, respectively, the start and end of e. Let us write $W = \sum_v W_v$, where W_v is in the linear span of closed paths that start at v. Thus for example $\cup_v = \sum_e \sqrt{\frac{\mu(v)}{\mu(w)}} X_e X_e^*$, where the sum is taken over all edges e starting at v.

With this notation, the expression

$$d\nu_N = Z_N^{-1} \exp\left(-N \sum_v \mu(v) Tr(\cup_v)\right) \prod_e dX_e$$

makes sense and gives us a probability measure, with respect to which we can choose our random matrix ensemble $\{X_e\}$.

For any $Q \in P_k$, the expression

$$\tau_N(Q) = \int \sum_v \frac{\mu(v)}{N} Tr(P(Q_v(X_e : e \in \Gamma)))d\nu_N$$

gives rise to a non-commutative law on the non-commutative probability space $(\mathcal{P}^\Gamma, \wedge_0)$ and so in particular on $(\mathcal{P}, \wedge_0)$. We denote this restriction by $\tau^{(N)}$.

Theorem 4.5. *With the above notation, $\tau^{(N)} \to \tau_{TL}$, where τ_{TL} is the Voiculescu trace on the planar $\mathcal{P}$.*

4.6. Random matrix ensembles. More generally, let us assume that we are given a non-commutative polynomial $V(t_1, \ldots, t_K, t_1^*, \ldots, t_K^*)$ which is a sum of monomials of the form $t_{i_1} t_{j_1}^* \cdots t_{i_p} t_{j_p}^*$. Then consider on $(M_{N\times N})^K$ the measure

$$d\mu_V^{(N)}(A_1, \ldots, A_K, A_1^*, \ldots, A_K^*) = \frac{1}{Z_N} 1_{\{\|A_j\| \le R\}} \exp(-NTr(V(A_1, \ldots, A_K, A_1^*, \ldots, A_K^*))) dA_1 \cdots dA_K dA_1^* \cdots dA_K^*, \quad (4.6.1)$$

where dA_j stands for Lebesgue measure on the j-th copy of $M_{N\times N}$. The constant Z_N is chosen so that $\mu_V^{(N)}$ is a probability measure (the cutoff R insures that the support of $\mu_V^{(N)}$ is compact). Of course, $R = \infty$ and $V(A_1, \ldots, A_K) = \sum A_k A_k^*$ corresponds to the Gaussian measure.

The measures $\mu_V^{(N)}$ are matrix analogs of the classical Gibbs measures $\mu_V = Z^{-1} \exp(-V(x))dx$.

Let us call the K-tuple of random matrices chosen from $(M_{N\times N}^{sa})^K$ at random according to this measure a *random multi-matrix ensemble* (see [AGZ10, Chapter 5]).

Certain properties of the random multi-matrix ensemble $A_1, \ldots, A_K$ is captured by the non-commutative laws $\tau_V^{(N)}$ defined on the algebra of non-commutative polynomials in $X_1, \ldots, X_K, X_1^*, \ldots, X_K^*$ by

$$\tau_V^{(N)}(Q(X_1, \ldots, X_K, X_1^*, \ldots, X_K^*)) = \int \frac{1}{N} Tr(Q(A_1, \ldots A_K, A_1^*, \ldots, A_K^*)) d\mu_V^{(N)}(A_1, \ldots, A_K, A_1^*, \ldots, A_K^*).$$

4.7. Combinatorial properties of the laws $\tau_V^{(N)}$. Remarkably, the laws $\tau_V^{(N)}$ have a very nice combinatorial interpretation. Let $P, W_1, \ldots, W_n$ be a monomials, and set $V(t_1, \ldots, t_K) = (\sum t_j t_j^*) + \sum_{j=1}^n \beta_j W_j$. Define a non-commutative law τ_V by

$$\tau_V(P) = \sum_{m_1, \ldots, m_n \geq 0} \sum_D \prod_{j=1}^n \frac{(-\beta_j)^{m_j}}{m_j!} \tag{4.7.1}$$

where the summation is taken over all planar tangles D with output disk labeled by P and having m_j interior disks labeled by W_j as in §3.6.1.

Theorem 4.8. *[Gui06, GMS06] Let $P, W_1, \ldots, W_n$ be monomials, and assume that $V(t_1, \ldots, t_K) = (\sum t_j t_j^*) + \sum_{j=1}^n \beta_j W_j$. Then for sufficiently small β_j,*

$$\tau_V^{(N)}(P) = \tau_V(P) + O(N^{-2}).$$

The right-hand side of (4.7.1) would make sense if we were to replace P and W_j by arbitrary elements of an arbitrary planar algebra (in fact, as written, equation (4.7.1) can be taken to occur in the planar algebra of polynomials). The term $\sum t_j t_j^*$ correpsonds to the element $\cup$ defined in Figure 3(b). We thus make the following definition.

Definition 4.9. Let $\mathcal{P}$ be a planar algebra, and assume that $Q \in P_k, W_j \in P_{k_j}$, $j = 1, \ldots, n$ are elements of algebra $\mathcal{P}$. Let $V_\beta = \cup + \sum_j \beta_j W_j$. We define the associated *free Gibbs law with symmetry* $\mathcal{P}$ to be the planar algebra law

$$\tau_{V_\beta}(Q) = \sum_{m_1, \ldots, m_n \geq 0} \sum_D \prod_{j=1}^n \frac{(-\beta_j)^{m_j}}{m_j!} Op(D)(P, \underbrace{W_1, \ldots, W_1}_{m_1}, \ldots, \underbrace{W_n, \ldots, W_n}_{m_n}). \tag{4.9.1}$$

Here the summation takes place over all planar tangles D having one disk of size k, m_1 input disks of size k_1, m_2 disks of size k_2, etc. and no output disks.

One can check that in the case of the planar algebra of polynomials, (4.9.1) is equivalent to (4.7.1).

Theorem 4.10. *Assume that $Q \in P_k, W_j \in P_{k_j}$, $j = 1, \ldots, n$ are elements of a finite-depth planar algebra $\mathcal{P}$, and let $V_\beta = \cup + \sum_j \beta_j W_j$. Then for sufficiently small β, the free Gibbs law given by* (4.9.1) *defines a non-negative trace on $(\oplus_{k\geq 0} P_k, \wedge_0)$.*

We now show that the laws τ_{V_β} arise from random matrix ensembles, just as in §4.4.1 (which corresponds to $\beta = 0$). Once again, we embed $\mathcal{P}$ into a graph planar algebra $\mathcal{P}^\Gamma$ and consider a family of random matrices X_e of size $[N\mu(v)] \times [N\mu(w)]$ labeled by the edges e of Γ (here $[\cdot]$ denotes the integer part of a number and μ is the Perron-Frobenius eigenvector of Γ). The matrices X_e are chosen according to the measure

$$d\nu_N = Z_N^{-1} \exp\left(-N \sum_v \mu(v) Tr\left((V_\beta)_v\right)\right) \prod_e dX_e.$$

For any $Q \in P_k$, the expression

$$\tau_N(Q) = \int \sum_v \frac{\mu(v)}{N} Tr(P(Q_v(X_e : e \in \Gamma))) d\nu_N$$

gives rise to a non-commutative law on the non-commutative probability space $(\mathcal{P}^\Gamma, \wedge_0)$ and, by restriction, on $(\mathcal{P}, \wedge_0)$. We denote this restriction by $\tau_{V_\beta}^{(N)}$.

Theorem 4.11. *Assume that $V = \cup + \sum_j \beta_j W_j$ as above. Then there is a $R_0 > 0$ so that for any $R > R_0$, there is a $\beta_0 > 0$ so that for all $|\beta_j| < \beta_0$, $\tau_V^{(N)} \to \tau_V$ where τ_V is as in Theorem 4.10.*

The finite-depth assumption seems to be technical in nature and is probably not necessary; it is automatically satisfied if $\delta < 2$.

4.12. Example: $O(n)$ models. One application of our construction sheds some light on the construction of so-called $O(n)$ models used by in physics by Zinn-Justin and Zuber in conjunctions with questions of knot combinatorics [ZJ03, ZJZ02]. For n an integer, the $O(n)$ model is the random matrix ensemble corresponding to the measure

$$Z_N^{-1} \exp(-NTr(V(X_1, \ldots, X_n))) dX_1 \cdots dX_n dX_1^* \cdots dX_n^*$$

where V is a fourth-degree polynomial in $X_1, \ldots, X_n, X_1^*, \ldots, X_n^*$, which is invariant under the $U(n)$ action given by (3.1.1). In degree ≤ 4, up to cyclic symmetry, the only such invariant polynomials actually lie in the copy of TL contained in the algebra $\mathcal{P}^{U(n)}$ in the notation of section §3.6.1: they are linear combinations of the constant polynomial and the polynomials $\cup = \sum X_i X_i^*$, $\cup\cup = \sum X_i X_i^* X_j X_j^*$ and $\Cup = \sum X_i X_j^* X_j X_i^*$ (these diagrams are in $TL \subset \mathcal{P}^{U(n)}$ with parameter $\delta = n$).

Hence the $O(n)$ model is the random matrix ensemble associated to the measure

$$\mu^{(N)}_{(\beta,n)} = Z_N^{-1} \exp\left(-NTr\left(\sum X_i X_i^* + \beta_i \sum X_i X_i^* X_j X_j^* + \beta_2 \sum X_i X_j^* X_j X_i^*\right)\right).$$

Thus we are led to consider the laws τ_β associated to the element

$$V_{(\beta,\delta)} = \cup + \beta_1 \cup^2 + \beta_2 \Cup \in TL$$

$\beta = (\beta_1, \beta_2)$ for each of the possible parameters $\delta \in \{2\cos\frac{\pi}{n} : n \geq 3\} \cup [2, +\infty)$. From our discussion we conclude that the limit law associated to the $O(n)$ model is exactly $\tau_{V_{(\beta,\delta=n)}}$.

But since our setting permits non-integer δ, we thus gain the flexibility of considering the laws $\tau_{V(\beta,\delta)}$ for other values of δ. It can be shown that the values of $\tau_{V_{(\beta,\delta=n)}}$ on a fixed element of TL are analytic in δ. Thus the extension we get is exactly the analytic extension from $n \in \mathbb{Z}$ to $\mathbb{C}$ considered by physicists in their analysis.

The combinatorics of the resulting law τ_V is governed by equation (4.9.1), which is written entirely in planar algebra terms. In particular, this shows that the $O(n)$ makes rigorous sense for any $\delta \in \{2\cos\frac{\pi}{n} : n \geq 3\} \cup [2, +\infty)$ (in the physics literature, the $O(n)$ model was used for non-integer n; the definition involved extending various equations analytically from $n \in \mathbb{Z}$ to $\mathbb{C}$).

It should be mentioned that $O(n)$ models were introduced in the physics literature to handle questions of knot enumerations; planar algebra interpretations of these computations are the subject of on-going research.

4.13. Properties of the limit laws τ_V. Because of Theorem 4.11, fixing a finite-depth planar algebra $\mathcal{P}$ and a family of elements $V_\beta = \cup + \beta W \in \mathcal{P}$, we obtain a family laws $\tau_\beta = \tau_{V_\beta}$. These in turn give rise to a family of von Neumann algebras $W^*(\tau_\beta)$ generated in the GNS representation associated to τ_β. When $\beta = 0$ these are free group factors (see Theorem 3.15). Voiculescu conjectured that this is also the case for $\beta \neq 0$ sufficiently small.

Using ideas from free probability theory, there has been significant progress on identifying properties of the associated Neumann algebras and C^*-algebras. The key is the following approximation result, whose proof relies on the theory of free stochastic differential equations [BS98].

Proposition 4.14. *[GS09] Assume that $\mathcal{P}$ is a the planar algebra of polynomials in K variables. Let $S_1, S_2, \ldots$ be an infinite free semicircular family generating the C^* algebra B with semicircular law τ, and let $A_\beta = C^*(\tau_\beta)$ in the GNS representation associated to τ_β. Let $X_1, \ldots, X_r \in A_\beta$. Then there is a $\beta_0 > 0$ so that for all $|\beta| < \beta_0$ and any $\epsilon > 0$ there exists an embedding $\alpha : A_\beta \to (A_\beta, \tau_\beta) * (B, \tau)$ and elements $Y_1, \ldots, Y_r \in B$ so that $\|\alpha(X_j) - Y_j\| < \epsilon$.*

Using this Proposition, many of the properties of the algebras A_β can be deduced from those of the algebra B.

Theorem 4.15. *[GS09] Let $V = \cup + \beta W$ be an element of a finite-depth planar algebra $\mathcal{P}$. Let τ_β be the associated law on $(\mathcal{P}, \wedge_0)$. The von Neumann algebra $M = W^*(\tau_\beta)$ and the C^*-algebra $A = C^*(\tau_\beta)$ satisfy:*

1. *M is a non-Γ II_1 factor and has the Haagerup property;*
2. *A is exact;*
3. *M has Ozawa's property AO and is therefore solid [Ash09].*

In the case that V is a polynomial potential (i.e., we are in the setting of Theorem 4.8), one can use the results of [PV82] to prove that $K_0(A) = 0$ and that A_β is projectionless. Indeed, if $p \in A$ were a non-trivial idempotent, then because of Proposition 4.14, $C^*(S_1, S_2, \dots) \subset C^*_{\text{red}}(\mathbb{F}_2)$ would be forced to contain a non-trivial idempotent as well. This statement has random matrix consequences:

Corollary 4.16. *[GS09] Let $\mathcal{P}$ be the planar algebra of polynomials in K variables, $V = V_\beta = \cup + \beta W \in \mathcal{P}$, and let $\tau_\beta = \tau_{V_\beta}$ be as in Theorem 4.8. Let $Q = Q^* \in \mathcal{P}$ be arbitrary polynomial. Let $Q^{(N)} = Q(X_1, \dots, X_K)$ be the random matrix obtained by evaluating Q in the random matrices $(X_1, \dots, X_K)$ chosen according to the measure* (4.6.1)*. Let $\mu^{(N)}$ be the expected value of the spectral measure of Q. Then $\mu^{(N)} \to \mu$ where μ is a measure with connected support.*

Proof. Let $Q^{(\infty)}$ denote the element of $C^*(\tau_\beta)$ that corresponds to the polynomial Q in the GNS construction associated to τ_β. Then the law of Q is exactly μ. If the support of μ is not connected, the spectrum of $Q \in C^*(\tau_\beta)$ is disconnected. But that means that $C^*(\tau_\beta)$ contains a non-trivial projection, contradicting Theorem 4.15. □

It turns out that in the presence of symmetry (for non-integer δ) the algebra A_β may contain non-trivial projections (even at $\beta = 0$). This phenomenon is not well-understood at this point, however. It would be interesting to compute the K-theory of the algebras A_β for general planar algebras $\mathcal{P}$.

References

[AGZ10] G. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Cambridge University Press, 2010.

[Ash09] J. Asher, *Free diffusions and property AO*, Preprint, arXiv.org/0907.1314, 2009.

[BIPZ78] E. Brézin, C. Itzykson, G. Parisi, and J. B. Zuber, *Planar diagrams*, Comm. Math. Phys. **59** (1978), no. 1, 35–51. MR MR0471676 (57 #11401)

[Bis97] Dietmar Bisch, *Bimodules, higher relative commutants and the fusion algebra associated to a subfactor*, Operator algebras and their applications (Waterloo, ON, 1994/1995), Fields Inst. Commun., vol. 13, Amer. Math. Soc., Providence, RI, 1997, pp. 13–63. MR MR1424954 (97i:46109)

[BS98] P. Biane and R. Speicher, *Stochastic calculus with respect to free Brownian motion and analysis on Wigner space*, Probab. Theory Related Fields **112** (1998), no. 3, 373–409. MR MR1660906 (99i:60108)

[Con] A. Connes, *Correspondences*, unpublished notes.

[Con94] ———, *Noncommutative geometry*, Academic Press, 1994.

[Dyk94] K. Dykema, *Interpolated free group factors*, Pacific J. Math. **163** (1994), 123–135.

[EM03] N. M. Ercolani and K. D. T.-R. McLaughlin, *Asymptotics of the partition function for random matrices via Riemann-Hilbert techniques and applications to graphical enumeration*, Int. Math. Res. Not. (2003), no. 14, 755–820. MR MR1953782 (2005f:82048)

[EZJ92] B. Eynard and J. Zinn-Justin, *The $O(n)$ model on a random surface: critical points and large-order behavior*, Nucl. Phys. B **386** (1992), 558–591.

[GHJ89] F.M. Goodman, R. de la Harpe, and V.F.R. Jones, *Coxeter graphs and towers of algebras*, Springer-Verlag, 1989.

[GJS08] A. Guionnet, V. Jones, and D. Shlyakhtenko, *Random matrices, free probability, planar algebras and subfactors*, To appear in Proceedings NCG, arXiv.org/0712.2904, 2008.

[GJS09] ———, *A semi-finite algebra associated to a planar algebra*. Preprint arXiv.org/0911.4728, 2009.

[GMS06] A. Guionnet and E. Maurel-Segala, *Combinatorial aspects of matrix models*, ALEA Lat. Am. J. Probab. Math. Stat. **1** (2006), 241–279.

[GS09] A. Guionnet and D. Shlyakhtenko, *Free diffusions and matrix models with strictly convex interaction*, Geom. Funct. Anal. **18** (2009), 1875–1916.

[Gui06] A. Guionnet, *Random matrices and enumeration of maps*, Proceedings Int. Cong. Math. **3** (2006), 623–636.

[Jon83] V.F.R. Jones, *Index for subfactors*, Invent. Math **72** (1983), 1–25.

[Jon99] ———, *Planar algebras*, Preprint, Berkeley, 1999.

[Jon01] ———, *The planar algebra of a bipartite graph*, Knots in the Hellas '98 (Delphi), World Scientific Publishing Co. Pte. Ltd., Singapore, 2001, pp. 94–117.

[JSW08] V.F.R. Jones, D. Shlyakhtenko, and K. Walker, *An orthogonal approach to the subfactor of a planar algebra*, to appear in Pacific J. Math, Preprint arXiv.org/0807.4146, 2008.

[Kos89] I. Kostov, *$O(n)$ vector model on a planar random lattice: spectrum of anomalous dimensions*, Modern Phys. Lett. A **4** (1989), 217–226.

[KS09a] V. Kodiyalam and V. S. Sunder, *Guionnet-Jones-Shlyakhtenko subfactors associated to finite-dimensional Kac algebras*, Preprint, arXiv.org:0901.3180, 2009.

[KS09b] ———, *On the Guionnet-Jones-Shlyakhtenko construction for graphs*, Preprint arXiv.org:0911.2047, 2009.

[MP67] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues in certain sets of random matrices*, Mat. Sb. (N.S.) **72 (114)** (1967), 507–536. MR MR0208649 (34 #8458)

[Pop86] S. Popa, *Correspondences*, INCREST preprint, 1986.

[Pop93] ———, *Markov traces on universal Jones algebras and subfactors of finite index*, Invent. Math. **111** (1993), 375–405.

[Pop95] ———, *An axiomatization of the lattice of higher relative commutants of a subfactor*, Invent. Math. **120** (1995), no. 3, 427–445. MR 96g:46051

[Pop99] ———, *Some properties of the symmetric enveloping algebras with applications to amenability and property T*, Documenta Mathematica **4** (1999), 665–744.

[Pop02] ———, *Universal construction of subfactors*, J. Reine Angew. Math. **543** (2002), 39–81. MR MR1887878 (2002k:46163)

[PS03] S. Popa and D. Shlyakhtenko, *Universal properties of $L(\mathbf{F}_\infty)$ in subfactor theory*, Acta Math. **191** (2003), no. 2, 225–257. MR MR2051399 (2005b:46140)

[PV82] M. Pimsner and D.-V. Voiculescu, *K-groups of reduced crossed products by free groups*, J. Operator Theory **8** (1982), 131–156.

[Răd94] F. Rădulescu, *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.

[Spe94] R. Speicher, *Multiplicative functions on the lattice of non-crossing partitions and free convolution*, Math. Annalen **298** (1994), 193–206.

[tH74] G. 't Hooft, *A planar diagram theory for strong interactions*, Nuclear Phys. B **72** (1974), 461–473.

[VDN92] D.-V. Voiculescu, K. Dykema, and A. Nica, *Free random variables*, CRM monograph series, vol. 1, American Mathematical Society, 1992.

[Voi85] D.-V. Voiculescu, *Symmetries of some reduced free product C^*-algebras*, Operator Algebras and Their Connections with Topology and Ergodic Theory, Lecture Notes in Mathematics, vol. 1132, Springer Verlag, 1985, pp. 556–588.

[Voi91] ———, *Limit laws for random matrices and free products*, Invent. math **104** (1991), 201–220.

[ZJ03] P. Zinn-Justin, *The general $\mathrm{O}(n)$ quartic matrix model and its application to counting tangles and links*, Comm. Math. Phys. **238** (2003), no. 1–2, 287–304. MR MR1990878 (2004d:57014)

[ZJZ02] P. Zinn-Justin and J.-B. Zuber, *Matrix integrals and the counting of tangles and links*, Discrete Math. **246** (2002), no. 1–3, 343–360, Formal power series and algebraic combinatorics (Barcelona, 1999). MR MR1887495 (2003i:57019)

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Rigidity for von Neumann Algebras and Their Invariants

Stefaan Vaes*

Abstract

We give a survey of recent classification results for von Neumann algebras $\mathrm{L}^\infty(X) \rtimes \Gamma$ arising from measure preserving group actions on probability spaces. This includes II_1 factors with uncountable fundamental groups and the construction of W^*-superrigid actions where $\mathrm{L}^\infty(X) \rtimes \Gamma$ entirely remembers the initial group action $\Gamma \curvearrowright X$.

Mathematics Subject Classification (2010). Primary 46L36; Secondary 46L40, 28D15, 37A20.

Keywords. Von Neumann algebra, II_1 factor, measure preserving group action, fundamental group of a II_1 factor, outer automorphism group, W^*-superrigidity.

1. Classifying II_1 Factors, a Panoramic Overview

A von Neumann algebra is a an algebra of bounded linear operators on a Hilbert space that is closed under the adjoint $*$-operation and that is closed in the weak operator topology. Von Neumann algebras arise naturally in the study of groups and their actions on measure spaces. These constructions go back to Murray and von Neumann's seminal papers [MvN36, Chapter XII] and [MvN43, §5.3].

- If Γ is a countable group, the left translation unitary operators on $\ell^2(\Gamma)$ generate the *group von Neumann algebra* $\mathrm{L}\Gamma$.

*Partially supported by ERC Starting Grant VNALG-200749, Research Programme G.0231.07 of the Research Foundation – Flanders (FWO) and K.U.Leuven BOF research grant OT/08/032.

K.U.Leuven, Department of Mathematics, Celestijnenlaan 200B, B-3001 Leuven (Belgium). E-mail: stefaan.vaes@wis.kuleuven.be.

- Every action $\Gamma \curvearrowright (X, \mu)$ of a countable group Γ by measurable transformations of a measure space (X, μ) and preserving sets of measure zero, gives rise to the *group measure space von Neumann algebra* $\mathrm{L}^\infty(X) \rtimes \Gamma$.

It is a central problem in the theory of von Neumann algebras to classify $\mathrm{L}\Gamma$ and $\mathrm{L}^\infty(X) \rtimes \Gamma$ in terms of the group Γ or the group action $\Gamma \curvearrowright (X, \mu)$. More generally, classifying or distinguishing families of von Neumann algebras is extremely challenging. In the first part of this exposition, I give a panoramic overview of the spectacular progress that has been made in this area over the last years. The overview is more thematically ordered than chronologically and necessarily incomplete. Several related important topics, including Jones' theory of subfactors or Voiculescu's free probability theory, are not treated.

All notions that are written in italics are defined in the preliminary section 2.

1.1. II_1 factors. The 'simple' von Neumann algebras M are those that cannot be written as a direct sum of two. Equivalently, the center of M is trivial and M is called a factor. Murray and von Neumann have classified factors into three types [MvN36] and proven that every von Neumann algebra can be decomposed as a direct integral of factors [vN49]. Connes [Co72] showed how general factors can be built up from those that admit a finite positive *trace,* called *II_1 factors.* The final form of this decomposition theory is due to Connes and Takesaki [Ta73, CT76]. Altogether, II_1 factors form the basic building blocks of arbitrary von Neumann algebras.

The *group von Neumann algebras* $\mathrm{L}\Gamma$ always admit a finite positive trace and are factorial if and only if Γ has infinite conjugacy classes (icc). When $\Gamma \curvearrowright (X, \mu)$ is essentially free, ergodic[1] and probability measure preserving (p.m.p.), the *group measure space von Neumann algebra* $\mathrm{L}^\infty(X) \rtimes \Gamma$ is a II_1 factor. Moreover, as proven by Singer [Si55], its isomorphism class only depends on the *equivalence relation* given by the orbits of $\Gamma \curvearrowright (X, \mu)$. This lead to the study of group actions up to orbit equivalence [Dy58] and we refer to [Sh05, Po06b, Fu09, Ga10] for surveys of the recent developments in this area of ergodic theory.

1.2. (Non)-isomorphism of II_1 factors. Two von Neumann algebras can be isomorphic in unexpected ways. While all hyperfinite[2] II_1 factors were already shown to be isomorphic in [MvN43], the culmination came with Connes' uniqueness theorem for *amenable* II_1 factors [Co75b] implying that all $\mathrm{L}\Gamma$ and all $\mathrm{L}^\infty(X) \rtimes \Gamma$ are isomorphic when Γ is any amenable icc group or $\Gamma \curvearrowright (X, \mu)$ is an arbitrary free ergodic p.m.p. action of an amenable group.

[1] Essential freeness means that for every $g \neq e$, the set of $x \in X$ with $g \cdot x = x$ has measure zero. Ergodicity means that globally Γ-invariant measurable subsets have either measure 0 or a complement of measure 0.

[2] A von Neumann algebra is hyperfinite if it is the direct limit of finite dimensional subalgebras.

In the early years examples of non-isomorphic II_1 factors M were obtained by analyzing asymptotically central sequences[3] of elements in M: property Gamma[4] [MvN43] allowed to prove that the free group factors $\mathrm{L}\mathbb{F}_n$ are not hyperfinite, a refinement yielded uncountably many non-isomorphic II_1 factors [McD69] and the χ-invariant [Co75a] provided the first examples where M is non-isomorphic to its opposite algebra M^{op} and where $M \not\cong M\overline{\otimes}M$.

The first rigidity phenomena for von Neumann algebras were discovered by Connes [Co80] who showed that the *fundamental group*[5] of $\mathrm{L}\Gamma$ is countable when Γ is an icc property (T) group. Several properties of groups – including *property (T),* the Haagerup property and related approximation properties – were shown in [CJ83, CH88, Jo00] to actually be properties of $\mathrm{L}\Gamma$, leading to remarkable non-isomorphism and non-embeddability theorems for group von Neumann algebras. Altogether it became clear that the world of II_1 factors is extremely rich, but that understanding the natural examples $\mathrm{L}\Gamma$ or $\mathrm{L}^\infty(X) \rtimes \Gamma$ in terms of the initial group or action is intrinsically very difficult.

1.3. Popa's deformation/rigidity theory. A major breakthrough in the classification of II_1 factors was realized by Popa and his discovery of deformation/ rigidity theory [Po01] (see [Po06b, Va06a] for a survey). Typically, Popa studies von Neumann algebras M that have a rigid subalgebra – e.g. given by the *relative property (T)* – such that the 'complement' has a strong deformation property. This gives the rigid subalgebra a canonical position within the ambient von Neumann algebra and has lead in [Po01] to the first example of a II_1 factor with trivial *fundamental group:* $M = \mathrm{L}(\mathbb{Z}^2 \rtimes \mathrm{SL}(2,\mathbb{Z})) = \mathrm{L}^\infty(\mathbb{T}^2) \rtimes \mathrm{SL}(2,\mathbb{Z})$. The canonical position of $\mathrm{L}^\infty(\mathbb{T}^2)$ implies that the fundamental group of M equals the fundamental group of the *orbit equivalence relation* of the action $\mathrm{SL}(2,\mathbb{Z}) \curvearrowright \mathbb{T}^2$, which is trivial because of [Ga99, Ga01].

In [Po03, Po04] Popa established a striking progress in his deformation/rigidity program by proving the following strong rigidity theorem for group measure space factors. Take an arbitrary free ergodic p.m.p. action $\Gamma \curvearrowright (X,\mu)$ of a property (T) group Γ and let $\Lambda \curvearrowright (Y,\eta) := (Y_0,\eta_0)^\Lambda$ be the *Bernoulli action* of an arbitrary icc group Λ. If the group measure space factors $\mathrm{L}^\infty(X) \rtimes \Gamma$ and $\mathrm{L}^\infty(Y) \rtimes \Lambda$ are isomorphic, then Γ must be isomorphic with Λ and their actions must be conjugate. Popa's strong rigidity theorem was the first result ever where conjugacy of actions could be deduced from the mere isomorphism of group measure space factors.

[3] A bounded sequence (x_n) in a II_1 factor M is called asymptotically central if $x_n y - y x_n$ converges to 0 in the strong operator topology for every $y \in M$.

[4] A II_1 factor has property Gamma if it admits an asymptotically central sequence of unitaries having trace 0.

[5] The fundamental group $\mathcal{F}(M)$ of a II_1 factor M consists of the numbers $\tau(p)/\tau(q)$ where p and q run over the projections in M satisfying $pMp \cong qMq$. Here τ denotes the trace on M.

1.4. Fundamental groups of II_1 factors. Progress in the classification of group measure space factors went hand in hand with major developments in the calculation of invariants of II_1 factors. The most well known invariant of M is the fundamental group $\mathcal{F}(M)$ of Murray and von Neumann [MvN43]. In one of their long-standing questions they asked what subgroups of $\mathbb{R}_+^*$ might occur as $\mathcal{F}(M)$.

Until 10 years ago, progress on this question has been scarce. Some II_1 factors, including the hyperfinite II_1 factor [MvN43] and $\mathrm{L}\mathbb{F}_\infty$ [Vo89, Ra91], were shown to have fundamental group $\mathbb{R}_+^*$ while Connes [Co80] proved that $\mathcal{F}(\mathrm{L}\Gamma)$ is countable whenever Γ is an icc property (T) group. A breakthrough in the understanding of fundamental groups came with Popa's first examples of II_1 factors having trivial fundamental group [Po01] and having prescribed countable fundamental group [Po03]. It remained a major open problem whether uncountable groups $\neq \mathbb{R}_+^*$ could appear as fundamental group.

In [PV08a] we solved this problem and proved that $\mathcal{F}(\mathrm{L}^\infty(X) \rtimes \mathbb{F}_\infty)$ ranges over a large family of subgroups of $\mathbb{R}_+^*$, including all countable subgroups and many uncountable subgroups that can have arbitrary Hausdorff dimension between 0 and 1. A similar result is true [PV08c] when $\mathbb{F}_\infty$ is replaced by almost any infinite free product of non-trivial groups, while $\mathcal{F}(\mathrm{L}^\infty(X) \rtimes \Gamma)$ is necessarily trivial when $\Gamma \curvearrowright (X, \mu)$ is an arbitrary free ergodic p.m.p. action of a free product of two finitely generated groups, one of them having property (T). So far, [PV08a, PV08c] provide the only known constructions of group measure space factors with fundamental group different from $\{1\}$ or $\mathbb{R}_+^*$.

The fundamental group $\mathcal{F}(M)$ of a II_1 factor M can also be viewed as the set of $t > 0$ such that the II_∞ factor $M\overline{\otimes}\mathrm{B}(H)$ admits an automorphism scaling the (infinite) trace $\tau \otimes \mathrm{Tr}$ by t. In [PV08a] we provide examples where $\mathcal{F}(M) = \mathbb{R}_+^*$, although $M\overline{\otimes}\mathrm{B}(H)$ admits no continuous trace-scaling action of $\mathbb{R}_+^*$.

1.5. Outer automorphisms and generalized symmetries. Another invariant of a II_1 factor M is its outer automorphism group[6] $\mathrm{Out}\, M$. In [IPP05] Ioana, Peterson and Popa established Bass-Serre isomorphism and subgroup (rather subalgebra) theorems for amalgamated free products of von Neumann algebras. As a consequence they obtained the first calculations of outer automorphism groups and proved that $\mathrm{Out}\, M$ can be any compact abelian group. In particular, they positively answered the question on the existence of II_1 factors without outer automorphisms. Later we showed [FV07] that in fact $\mathrm{Out}\, M$ can be any compact group. The results in [IPP05, FV07] are existence theorems involving a Baire category argument. We obtained the first concrete and

[6]The outer automorphism group $\mathrm{Out}\, M$ is defined as the quotient $\mathrm{Aut}\, M/\mathrm{Inn}\, M$, where $\mathrm{Inn}\, M$ denotes the normal subgroup of $\mathrm{Aut}\, M$ consisting of the inner automorphisms $\mathrm{Ad}\, u$, $u \in \mathcal{U}(M)$.

explicit calculations of $\mathrm{Out}\, M$ in [PV06, Va07] and proved that $\mathrm{Out}\, M$ can be any countable group.

Both the elements of the fundamental group and the automorphisms of a II_1 factor M give rise to *Hilbert M-M-bimodules* ${}_M\mathcal{H}_M$. An M-M-bimodule that is finitely generated, both as a left and as a right M-module is said to be of finite Jones index [Jo82]. The finite index M-M-bimodules form a C*-tensor category $\mathrm{Bimod}\, M$ and this should be considered as the generalized (or quantum) symmetry group of M. Whenever $M \subset P$ is a finite index subfactor, ${}_M\mathrm{L}^2(P)_M$ is a finite index bimodule. In this sense, $\mathrm{Bimod}\, M$ also encodes the subfactor structure of M. In [Va06b] I proved the existence of II_1 factors M such that $\mathrm{Bimod}\, M$ is trivial, i.e. only consists of multiples of the trivial bimodule $\mathrm{L}^2(M)$. Such II_1 factors have trivial fundamental group, trivial outer automorphism group and no non-trivial finite index subfactors. Explicit examples were provided in [Va07] where also several concrete calculations of $\mathrm{Bimod}\, M$ were made. These calculations were exploited in [DV10] to give a full classification of all finite index subfactors of certain II_1 factors. In [FV08] every representation category of a compact group K is realized as $\mathrm{Bimod}\, M$. More precisely, for every compact group K we prove the existence of a minimal action of K on a II_1 factor M such that, denoting by M^K the subfactor of K-invariant elements, the natural faithful tensor functor $\mathrm{Rep}\, K \to \mathrm{Bimod}(M^K)$ is 'surjective', i.e. an equivalence of categories.

1.6. W*-superrigidity and uniqueness of Cartan subalgebras. Two free ergodic p.m.p. actions $\Gamma \curvearrowright (X, \mu)$ and $\Lambda \curvearrowright (Y, \eta)$ are called

- conjugate (or isomorphic), if there exists an isomorphism of probability spaces $\Delta : X \to Y$ and an isomorphism of groups $\delta : \Gamma \to \Lambda$ satisfying $\Delta(g \cdot x) = \delta(g) \cdot \Delta(x)$ for all $g \in \Gamma$ and a.e. $x \in X$;
- orbit equivalent (OE), if there exists an isomorphism of probability spaces $\Delta : X \to Y$ satisfying $\Delta(\Gamma \cdot x) = \Lambda \cdot \Delta(x)$ for a.e. $x \in X$;
- W*-equivalent, if $\mathrm{L}^\infty(X) \rtimes \Gamma \cong \mathrm{L}^\infty(Y) \rtimes \Lambda$.

Obviously, conjugacy of actions implies orbit equivalence and Singer [Si55] proved that an orbit equivalence is the same as a W*-equivalence sending the *group measure space Cartan subalgebras*[7] $\mathrm{L}^\infty(X)$ and $\mathrm{L}^\infty(Y)$ onto each other. Rigidity theory for group actions aims at establishing the converse implications under appropriate assumptions. Pioneering OE rigidity results were obtained

[7] In general, a Cartan subalgebra A of a II_1 factor M is a maximal abelian subalgebra whose normalizing unitaries generate M. Whenever $\Gamma \curvearrowright (X, \mu)$ is a free ergodic p.m.p. action, $\mathrm{L}^\infty(X) \subset \mathrm{L}^\infty(X) \rtimes \Gamma$ is an example of a Cartan subalgebra that we call of group measure space type. Not all II_1 equivalence relations can be implemented by a free action of a countable group [Fu98b] and hence a general Cartan subalgebra need not be of group measure space type.

by Zimmer [Zi79, Zi84] and the first breakthrough W*-rigidity theorems were proven by Popa [Po01, Po03, Po04] (see paragraph 1.3). Our aim here however is to discuss the ideal kind of rigidity, labeled W*- (respectively OE-) superrigidity, where the entire isomorphism class of $\Gamma \curvearrowright (X, \mu)$ is recovered from its W*-class (resp. OE class).

While a striking number of OE superrigid actions have been discovered over the last 10 years [Fu98b, Po05, Po06a, Ki06, Io08, PV08b, Ki09, PS09], the first W*-superrigid actions were only discovered very recently in my joint paper with Popa [PV09]. We found a family of amalgamated free product groups Γ and a large class of W*-superrigid Γ-actions, including (generalized) Bernoulli actions, Gaussian actions and certain co-induced actions.

Note that W*-superrigidity for an action $\Gamma \curvearrowright (X, \mu)$ is equivalent to the 'sum' between its OE superrigidity and the uniqueness, up to unitary conjugacy, of $\mathrm{L}^\infty(X)$ as a group measure space Cartan subalgebra in $\mathrm{L}^\infty(X) \rtimes \Gamma$. This makes W*-superrigidity results extremely difficult to obtain, since each one of these problems is notoriously hard. Contrary to the long list of OE superrigid actions referred to in the previous paragraph, unique Cartan decomposition proved to be much more challenging to establish, and the only existing results cover very particular group actions. Thus, a first such result, obtained by Ozawa and Popa [OP07], shows that given any profinite action $\Gamma \curvearrowright X$ of a product of free groups $\Gamma = \mathbb{F}_{n_1} \times \cdots \times \mathbb{F}_{n_k}$, with $k \geq 1$, $2 \leq n_i \leq \infty$, all Cartan subalgebras of $M = \mathrm{L}^\infty(X) \rtimes \Gamma$ are unitarily conjugate to $\mathrm{L}^\infty(X)$. A similar result, covering groups Γ that have the complete metric approximation property and that admit a proper 1-cocycle into a non-amenable representation, was then proved in [OP08]. More recently, Peterson showed [Pe09] that factors arising from profinite actions of non-trivial free products $\Gamma = \Gamma_1 * \Gamma_2$, with at least one of the Γ_i not having the Haagerup property, have a unique group measure space Cartan subalgebra, up to unitary conjugacy. But so far, none of these group actions could be shown to be OE superrigid. Nevertheless, an intricate combination of results in [Io08, OP08, Pe09] were used to prove the existence of virtually[8] W*-superrigid group actions $\Gamma \curvearrowright X$ in [Pe09], by a Baire category argument.

In [PV09] we established a very general unique Cartan decomposition result, which allowed us to obtain a wide range of W*-superrigid group actions. Thus, we first proved the uniqueness, up to unitary conjugacy, of the group measure space Cartan subalgebra in the II_1 factor given by an arbitrary free ergodic p.m.p. action of any group Γ belonging to a large family $\mathcal{G}$ of amalgamated free product groups. By combining this with Kida's OE superrigidity in [Ki09], we deduced that if $T_n < \mathrm{PSL}(n, \mathbb{Z})$ denotes the group of upper triangular matrices in $\mathrm{PSL}(n, \mathbb{Z})$, then any free mixing p.m.p. action of $\Gamma = \mathrm{PSL}(n, \mathbb{Z}) *_{T_n} \mathrm{PSL}(n, \mathbb{Z})$

[8] Following [Fu98a], *virtual* means that the ensuing conjugacy of $\Gamma \curvearrowright X$ and $\Lambda \curvearrowright Y$ is up to finite index subgroups of Γ, Λ.

is W*-superrigid. In combination with [Po05, Po06a], we proved that for many groups Γ in the family $\mathcal{G}$, the Bernoulli actions of Γ are W*-superrigid.

Very recently, Ioana [Io10] obtained the beautiful result that all Bernoulli actions of property (T) groups are W*-superrigid.

Recall from paragraph 1.3 Popa's strong rigidity theorem for Bernoulli actions and note the asymmetry in the formulation: there is a (rigidity) condition on the group Γ and a (deformation) condition on the action $\Lambda \curvearrowright (Y, \eta)$. One of the novelties of [PV09] is a transfer of rigidity principle showing that under W*-equivalence of $\Gamma \curvearrowright X$ and $\Lambda \curvearrowright Y$, some of the rigidity properties of Γ persist in the arbitrary unknown group Λ. Note however that property (T) itself is not stable under W*-equivalence: there exist W*-equivalent group actions such that Γ has property (T) while Λ has not (see Section 5).

1.7. Indecomposability results. Related to the uniqueness problem of Cartan subalgebras obviously is the existence question. Voiculescu [Vo95] proved that the free group factors admits no Cartan subalgebra, because the presence of a Cartan subalgebra forces the free entropy dimension of any generating set to be smaller or equal than 1. Another application of Voiculescu's free entropy theory was given by Ge [Ge96] who showed that the free group factors are prime: they cannot be written as the tensor product of two II_1 factors.

Using delicate C*-algebra techniques, Ozawa [Oz03] proved that for all icc word hyperbolic groups Γ – in particular when Γ equals the free group $\mathbb{F}_n$ – the group factor $\mathrm{L}\Gamma$ is solid: the relative commutant $A' \cap \mathrm{L}\Gamma$ of an arbitrary diffuse[9] subalgebra is injective. Obviously non-hyperfinite solid II_1 factors, as well as all their non-hyperfinite subfactors, are prime. A combination of techniques from [Po01, Oz03] then allowed Ozawa and Popa [OP03] to introduce a family of II_1 factors that have an essentially unique tensor product decomposition into prime factors. Peterson's L^2-rigidity [Pe06] – a II_1 factor analogue for the vanishing of the first ℓ^2-Betti number of a group – as well as Bass-Serre rigidity for free products of von Neumann algebras [IPP05, CH08] provided further examples of prime factors.

The free group factors have no Cartan subalgebra and are solid. In [OP07] both properties are brought together and $\mathrm{L}\mathbb{F}_n$ is shown to be strongly solid: the normalizer of an arbitrary diffuse abelian subalgebra is hyperfinite. Other examples of strongly solid II_1 factors were given in [Ho09, HS09].

Organization of the paper. In so far as the above gave an overview of some recent developments, in the rest of the paper I present the main ideas behind a number of chosen topics: Popa's deformation/rigidity theory in Section 3, computations of fundamental groups in Section 4, the (non-)uniqueness of Cartan subalgebras in Section 5 and W*-superrigidity in Section 6.

[9] A von Neumann algebra is called diffuse if it admits no minimal projections.

2. Preliminaries

Traces, $\mathbf{II_1}$ factors and the Hilbert bimodule $\mathbf{L^2(M)}$. A *finite trace* on a von Neumann algebra M is a linear map $\tau : M \to \mathbb{C}$ satisfying $\tau(xy) = \tau(yx)$ for all $x, y \in M$. We say that τ is *positive* if $\tau(x) \geq 0$ for all positive operators $x \in M$. A positive trace τ is called *faithful* if the equality $\tau(x^*x) = 0$ implies that $x = 0$. A positive trace τ is called *normal* if τ is weakly continuous on the unit ball of M.

A *II_1 factor* is a von Neumann algebra with trivial center that admits a non-zero finite positive trace τ and that is non-isomorphic to a matrix algebra $\mathrm{M}_n(\mathbb{C})$. Normalizing τ such that $\tau(1) = 1$, the trace is unique. Moreover τ is automatically normal. We denote by $\|x\|_2 = \sqrt{\tau(x^*x)}$ the L^2-norm corresponding to τ. Completing M w.r.t. the scalar product $\langle x, y\rangle = \tau(x^*y)$ yields the Hilbert space $\mathrm{L}^2(M)$, which is an M-M-bimodule by left and right multiplication on M.

We denote by Tr the non-normalized trace on $\mathrm{M}_n(\mathbb{C})$. Occasionally, Tr denotes the infinite trace on positive operators in $\mathrm{B}(\mathcal{H})$.

Group von Neumann algebras. Let Γ be a countable group. Then $\mathrm{L}\Gamma$ is the unique tracial von Neumann algebra generated by unitary elements $(u_g)_{g\in\Gamma}$ with the following two properties: $u_g u_h = u_{gh}$ for all $g, h \in \Gamma$ and $\tau(u_g) = 0$ for all $g \neq e$. Alternatively, we denote by $(\delta_g)_{g\in\Gamma}$ the standard orthonormal basis of $\ell^2(\Gamma)$, define the translation unitary operators u_g as $u_g\delta_h = \delta_{gh}$ and define $\mathrm{L}\Gamma$ as the von Neumann algebra generated by $\{u_g \mid g \in \Gamma\}$, with τ being given by $\tau(x) = \langle \delta_e, x\delta_e\rangle$ for all $x \in \mathrm{L}\Gamma$.

Group measure space construction. If (P, τ) is a tracial von Neumann algebra and $\Gamma \overset{\alpha}{\curvearrowright} P$ is an action of a countable group Γ by trace preserving automorphisms $\alpha_g \in \operatorname{Aut} P$, the *crossed product* $P \rtimes \Gamma$ is the unique tracial von Neumann algebra (M, τ) generated by a trace-preserving copy of P and unitary elements $(u_g)_{g\in\Gamma}$ satisfying the following properties:

$$u_g a u_g^* = \alpha_g(a) \text{ for all } g \in \Gamma, a \in P\,, \quad u_g u_h = u_{gh} \text{ for all } g, h \in \Gamma\,,$$
$$\tau(au_g) = 0 \text{ for all } a \in P, g \neq e\,.$$

The map $au_g \mapsto a \otimes \delta_g$ provides an identification $\mathrm{L}^2(P \rtimes \Gamma) = \mathrm{L}^2(P)\overline{\otimes}\ell^2(\Gamma)$ and then an explicit realization of $P \rtimes \Gamma$ as an algebra of bounded operators on the Hilbert space $\mathrm{L}^2(P)\overline{\otimes}\ell^2(\Gamma)$.

When $\Gamma \curvearrowright (X, \mu)$ is a probability measure preserving (p.m.p.) action, one considers the corresponding trace preserving action $\Gamma \curvearrowright \mathrm{L}^\infty(X)$ and constructs $M = \mathrm{L}^\infty(X) \rtimes \Gamma$. The abelian subalgebra $\mathrm{L}^\infty(X) \subset M$ is maximal abelian if and only if $\Gamma \curvearrowright (X, \mu)$ is essentially free, meaning that for all $g \neq e$ the set $\{x \in X \mid g \cdot x = x\}$ has measure zero. When $\Gamma \curvearrowright (X, \mu)$ is essentially free, the center of M equals $\mathrm{L}^\infty(X)^\Gamma$, the algebra of Γ-invariant functions in $\mathrm{L}^\infty(X)$. Hence, factoriality of M is then equivalent with ergodicity of $\Gamma \curvearrowright (X, \mu)$.

(Generalized) Bernoulli actions. If Γ is an infinite countable group and if (X_0, μ_0) is a non-trivial probability space, define the infinite product $(X, \mu) := (X_0, \mu_0)^\Gamma$ on which Γ acts by shifting the indices: $(g \cdot x)_h = x_{g^{-1}h}$. The action $\Gamma \curvearrowright (X, \mu)$ is called the *Bernoulli action* with base space (X_0, μ_0) and it is a free ergodic p.m.p. action.

More generally, if Γ acts on the countably infinite set I, one considers the *generalized Bernoulli action* $\Gamma \curvearrowright (X_0, \mu_0)^I$ given by $(g \cdot x)_i = x_{g^{-1} \cdot i}$. This action is p.m.p. and it is ergodic if and only if every orbit $\Gamma \cdot i$ is infinite. If (X_0, μ_0) is non-atomic, essential freeness is equivalent with every $g \neq e$ acting non-trivially on I. If (X_0, μ_0) has atoms, essential freeness is equivalent with every $g \neq e$ moving infinitely many $i \in I$.

Bimodules. An M-N-bimodule ${}_M\mathcal{H}_N$ between von Neumann algebras M and N is a Hilbert space $\mathcal{H}$ equipped with a normal unital $*$-homomorphism $\lambda : M \to \mathrm{B}(\mathcal{H})$ and a normal unital $*$-anti-homomorphism $\rho : N \to \mathrm{B}(\mathcal{H})$ such that $\lambda(M)$ and $\rho(N)$ commute. We write $x\xi y$ instead of $\lambda(x)\rho(y)\xi$.

Bimodules should be considered as the II_1 factor analogue of unitary group representations. Based on this philosophy, several representation theoretic properties of groups have a II_1 factor counterpart: amenability, the Haagerup property, property (T), etc.

To establish the dictionary between group representations and bimodules, take a countable group Γ and put $M = \mathrm{L}\Gamma$. Whenever $\pi : \Gamma \to \mathcal{U}(\mathcal{K})$ is a unitary representation, define the Hilbert space $\mathcal{H}_\pi = \ell^2(\Gamma) \overline{\otimes} \mathcal{K}$ and turn $\mathcal{H}_\pi$ into an M-M-bimodule by putting $u_g(\delta_h \otimes \xi)u_k := \delta_{ghk} \otimes \pi(g)\xi$. The trivial representation corresponds to the *trivial bimodule* ${}_M\mathrm{L}^2(M)_M$, the regular representation corresponds to the *coarse bimodule* ${}_{M \otimes 1}\bigl(\mathrm{L}^2(M) \overline{\otimes} \mathrm{L}^2(M)\bigr)_{1 \otimes M}$ and one defines the notions of *containment* and *weak containment* of bimodules [Po86] in such a way that through the construction $\pi \rightsquigarrow \mathcal{H}_\pi$ these notions exactly correspond to the well known concepts from representation theory. Finally, the *Connes tensor product* of bimodules [Co94, V.Appendix B] is so that $\mathcal{H}_\pi \otimes_M \mathcal{H}_\rho \cong \mathcal{H}_{\pi \otimes \rho}$.

Definition 2.1. A tracial von Neumann algebra (M, τ) is called *amenable* [Po86] if the coarse bimodule weakly contains the trivial bimodule.

We say that (M, τ) has *property (T)* [CJ83] if any M-M-bimodule weakly containing the trivial bimodule, must contain the trivial bimodule.

We finally say that the subalgebra $N \subset M$ has the *relative property (T)* [Po01] if any M-M-bimodule weakly containing the trivial bimodule, must contain the bimodule ${}_N\mathrm{L}^2(M)_M$.

Completely positive maps and bimodules. A linear map $\varphi : M \to N$ is called *completely positive* if for every n the amplified map $\mathrm{id} \otimes \varphi : \mathrm{M}_n(\mathbb{C}) \otimes M \to \mathrm{M}_n(\mathbb{C}) \otimes N$ maps positive operators to positive operators. In the same way as unitary representations are related to positive definite functions, also bimodules and completely positive maps form two sides of the same story. Whenever $\varphi : M \to N$ is a unital trace preserving completely positive map,

the separation and completion of the algebraic tensor product $M \otimes_{\rm alg} N$ w.r.t. the scalar product $\langle a \otimes b, c \otimes d \rangle = \tau(b^* \varphi(a^* c) d)$ defines a Hilbert space $\mathcal{H}$ that naturally becomes an M-N-bimodule. By construction $\xi = 1 \otimes 1$ is a cyclic vector, meaning that $M\xi N$ is dense in $\mathcal{H}$, and is a trace vector, meaning that $\tau(a) = \langle \xi, a\xi \rangle$ and $\tau(b) = \langle \xi, \xi b \rangle$ for all $a \in M, b \in N$. Every bimodule with a cyclic trace vector arises in this way.

Cartan subalgebras and equivalence relations. Whenever $A \subset M$ is a von Neumann subalgebra, we denote by $\mathcal{N}_M(A) := \{u \in \mathcal{U}(M) \mid uAu^* = A\}$ the group of unitaries in M that normalize A. A *Cartan subalgebra* $A \subset M$ of a II_1 factor is a maximal abelian subalgebra such that $\mathcal{N}_M(A)$ generates M. Whenever $\Gamma \curvearrowright (X, \mu)$ is a free ergodic p.m.p. action, $\mathrm{L}^\infty(X) \subset \mathrm{L}^\infty(X) \rtimes \Gamma$ is a Cartan subalgebra. We call such Cartan subalgebras of *group measure space* type.

The relevance of Cartan subalgebras in the study of group measure space factors stems from the following theorem.

Theorem 2.2 (Singer [Si55]). *Let $\Gamma \curvearrowright (X, \mu)$ and $\Lambda \curvearrowright (Y, \eta)$ be free ergodic p.m.p. actions and denote $A = \mathrm{L}^\infty(X)$, $B = \mathrm{L}^\infty(Y)$. Assume that $\Delta : X \to Y$ is an isomorphism of probability spaces with corresponding trace preserving isomorphism $\Delta_* : A \to B : F \mapsto F \circ \Delta^{-1}$. Then, the following are equivalent.*

- *Δ is an orbit equivalence: for almost every $x \in X$, we have $\Delta(\Gamma \cdot x) = \Lambda \cdot \Delta(x)$.*
- *Δ_* extends to a $*$-isomorphism $A \rtimes \Gamma \to B \rtimes \Lambda$.*

A *II_1 equivalence relation* [FM75] on a standard probability space (X, μ) is an equivalence relation $\mathcal{R} \subset X \times X$ with *countable* equivalence classes such that $\mathcal{R}$ is a Borel subset of $X \times X$ and such that $\mathcal{R}$ is *ergodic* and *probability measure preserving*. Here $\mathcal{R}$ is called ergodic if every $\mathcal{R}$-saturated Borel set has measure 0 or 1, while $\mathcal{R}$ is said to be p.m.p. if every bimeasurable bijection $\varphi : X \to X$ with graph inside $\mathcal{R}$, preserves μ.

Whenever $\Gamma \curvearrowright (X, \mu)$ is an ergodic p.m.p. action, the *orbit equivalence relation* $\mathcal{R}(\Gamma \curvearrowright X)$ is of type II_1. Every II_1 equivalence relation is of this form, but it is not always possible to choose an essentially *free* action implementing $\mathcal{R}$ (see [Fu98b] and [PV08b, Section 7]).

A variant of the group measure space construction [FM75] allows to associate a II_1 factor $\mathrm{L}\mathcal{R}$ to any II_1 equivalence relation $\mathcal{R}$ on (X, μ). By construction $\mathrm{L}\mathcal{R}$ contains a copy of $\mathrm{L}^\infty(X)$ as a Cartan subalgebra. Every Cartan inclusion $A \subset M$ arises in this way, modulo the possible appearance of a scalar 2-cocycle on $\mathcal{R}$. When $\Gamma \curvearrowright (X, \mu)$ is a free ergodic p.m.p. action, we canonically have $\mathrm{L}^\infty(X) \rtimes \Gamma = \mathrm{L}(\mathcal{R}(\Gamma \curvearrowright X))$. It is however important to note that both the crossed product and the orbit equivalence relation make sense for non-free actions, but no longer yield isomorphic von Neumann algebras.

Fundamental groups of II_1 factors. When M is a II_1 factor and $t > 0$, one defines as follows the amplification M^t. For $0 < t \leq 1$, take a projection $p \in M$ with $\tau(p) = t$ and put $M^t := pMp$. For larger t, take an integer n, a projection $p \in \mathrm{M}_n(\mathbb{C}) \otimes M$ with $(\mathrm{Tr} \otimes \tau)(p) = t$ and put $M^t := p(\mathrm{M}_n(\mathbb{C}) \otimes M)p$. As such, M^t is well defined up to isomorphism. One proves that $(M^t)^s \cong M^{ts}$. The *fundamental group* $\mathcal{F}(M)$ is defined as the set of $t > 0$ such that $M^t \cong M$. The fundamental group of a II_1 equivalence relation $\mathcal{R}$ is defined in a similar way. If $M = \mathrm{L}\mathcal{R}$ denotes the associated II_1 factor, by construction $\mathcal{F}(\mathcal{R}) \subset \mathcal{F}(M)$, but this inclusion can be strict [Po06a, Section 6.1].

3. Popa's Deformation/Rigidity Theory

Popa's deformation/rigidity theory, initiated in [Po01], has revolutionized our understanding of II_1 factors. We explain in this section what kind of deformations Popa introduced and how they can be combined with the rigidity given by the relative property (T).

Definition 3.1. A *deformation of the identity* on a tracial von Neumann algebra (M, τ) is a sequence of normal completely positive maps $\varphi_n : M \to M$ that are unital, trace preserving and satisfy

$$\|\varphi_n(x) - x\|_2 \to 0 \quad \text{for all } \; x \in M \; .$$

Both group factors $\mathrm{L}\Gamma$ and crossed products $P \rtimes \Gamma$ admit natural deformations of the identity. Indeed, if $\varphi : \Gamma \to \mathbb{C}$ is a positive definite function, both

$$\begin{aligned} &\mathrm{L}\Gamma \to \mathrm{L}\Gamma : u_g \mapsto \varphi(g) u_g \quad \text{for all } \; g \in \Gamma \quad \text{and} \\ &P \rtimes \Gamma \to P \rtimes \Gamma : a u_g \mapsto \varphi(g) a u_g \quad \text{for all } \; a \in P, g \in \Gamma \end{aligned}$$

extend to normal completely positive maps on $\mathrm{L}\Gamma$, resp. $P \rtimes \Gamma$. If φ is normalized, i.e. $\varphi(e) = 1$, these maps are unital and trace preserving.

Example 3.2. If Γ has the Haagerup approximation property [Ha78], there exists a sequence $\varphi_n : \Gamma \to \mathbb{C}$ of positive definite functions converging pointwise to 1 and with $\varphi_n \in c_0(\Gamma)$ for every n. As we discuss below, the corresponding deformation of the identity of $P \rtimes \Gamma$ plays a crucial role in [Po01].

If $\Gamma = \Gamma_1 * \Gamma_2$ is a free product and $|g|$ denotes the natural word length of $g \in \Gamma$ w.r.t. this free product decomposition and if $0 < \rho < 1$, then the formula $\varphi_\rho(g) = \rho^{|g|}$ defines a positive definite function on Γ. If $\rho \to 1$, then $\varphi_\rho \to 1$ pointwise. The corresponding deformation of the identity is the starting point for [IPP05, PV09] and also this is discussed below.

A second family of deformations of the identity arises as follows. Let $\Gamma \overset{\alpha}{\curvearrowright} (P, \tau)$ be a trace preserving action. Assume that $\varphi_n : P \to P$ is a deformation of the identity such that $\varphi_n \circ \alpha_g = \alpha_g \circ \varphi_n$ for all $g \in \Gamma$, $n \in \mathbb{N}$. Then, the formula $a u_g \mapsto \varphi_n(a) u_g$ defines a deformation of the identity on $P \rtimes \Gamma$.

Definition 3.3 ([Po03]). A p.m.p. action $\Gamma \curvearrowright (X, \mu)$ is *malleable* if there exists a continuous family $(\alpha_t)_{t\in[0,1]}$ of p.m.p. transformations of $X \times X$ such that for all $t \in [0, 1]$, α_t commutes with the diagonal action $g \cdot (x, y) = (g \cdot x, g \cdot y)$ and such that $\alpha_0 = \mathrm{id}$ and $\alpha_1(x, y)$ is of the form $(\ldots, x)$.

The *Bernoulli action* $\Gamma \curvearrowright [0, 1]^\Gamma$ is malleable. It suffices to construct a continuous family $(\alpha_t^0)_{t\in[0,1]}$ of p.m.p. transformations of the square $[0, 1] \times [0, 1]$ such that $\alpha_0^0 = \mathrm{id}$ and $\alpha_1^0(x, y) = (\ldots, x)$. This can be done by 'rotating the square' counterclockwise over 90 degrees. Next, one identifies $[0, 1]^\Gamma \times [0, 1]^\Gamma = ([0, 1] \times [0, 1])^\Gamma$ and defines $(\alpha_t(x, y))_g = \alpha_t^0(x_g, y_g)$. By construction, α_t commutes with the diagonal Γ-action.

Other examples of malleable actions are *generalized Bernoulli actions* $\Gamma \curvearrowright [0, 1]^I$ given by an action $\Gamma \curvearrowright I$ or *Gaussian actions* given by an orthogonal representation of Γ on a real Hilbert space.

Example 3.4. If α_t is a malleable deformation, put $A = \mathrm{L}^\infty(X)$ and define the corresponding automorphisms α_t of $A\overline{\otimes}A = \mathrm{L}^\infty(X \times X)$ given by $\alpha_t(F) := F(\alpha_t^{-1}(\,\cdot\,))$. By definition $\alpha_0 = \mathrm{id}$ and $\alpha_1(a\otimes 1) = 1\otimes a$ for all $a \in A$. View $A \hookrightarrow A\overline{\otimes}A : a \mapsto a \otimes 1$ and denote by $E : A\overline{\otimes}A \to A$ the trace preserving conditional expectation (which corresponds to integration w.r.t. the second variable). The formula $\varphi_t : A \to A : \varphi_t(a) = E_A(\alpha_t(a\otimes 1))$ defines a continuous family of unital trace preserving completely positive maps with $\varphi_0 = \mathrm{id}$ and $\varphi_1(a) = \tau(a)1$. When $t \to 0$, we get a deformation of the identity on $\mathrm{L}^\infty(X) \rtimes \Gamma$ that is at the heart of [Po03, Po04].

A variant of the malleable deformation for Bernoulli actions is the *tensor length deformation* [Io06]. Indeed, given a base probability space (X_0, μ_0) and a countable set I, put $(X, \mu) = (X_0, \mu_0)^I$ and identify $A := \mathrm{L}^\infty(X)$ with the infinite tensor product $\overline{\otimes}_{i\in I}(A_0, \tau)$, where $A_0 := \mathrm{L}^\infty(X_0)$. We write $A = A_0^I$. Whenever $J \subset I$ is a subset, we view A_0^J as a subalgebra of A_0^I. We then define for every $0 < \rho < 1$,

$$\theta_\rho : A \to A : \theta_\rho(a) = \rho^n a \quad \text{when } \; a \in (A_0 \ominus \mathbb{C}1)^J \; \text{ and } \; |J| = n \; .$$

Then θ_ρ is a well defined unital trace preserving normal completely positive map and $\theta_\rho \to \mathrm{id}$ when $\rho \to 1$. By construction, θ_ρ commutes with the generalized Bernoulli action $\Gamma \curvearrowright A_0^I$ whenever $\Gamma \curvearrowright I$.

Combining deformation and rigidity. Let $N \subset M$ be an inclusion with the relative property (T), see Definition 2.1. Whenever $\varphi_n : M \to M$ is a deformation of the identity, it follows that φ_n converges uniformly in $\|\cdot\|_2$ on the unit ball of N. We illustrate this combination of deformation and rigidity by indicating the main ideas behind two of Popa's theorems.

Theorem 3.5 (Popa [Po01]). *The II_1 factor $M = \mathrm{L}(\mathbb{Z}^2 \rtimes \mathrm{SL}(2, \mathbb{Z}))$ has trivial fundamental group.*

Write $A = \mathrm{L}^\infty(\mathbb{T}^2)$, $\Gamma = \mathrm{SL}(2,\mathbb{Z})$ and identify $M = A \rtimes \Gamma$. The group Γ has the Haagerup property and hence admits a sequence of positive definite functions φ_n such that $\varphi_n \in c_0(\Gamma)$ for all n and $\varphi_n \to 1$ pointwise. As in Example 3.2, we get a deformation of the identity on M given by $\theta_n(au_g) = \varphi_n(g)au_g$. Assume that $N \subset M$ has the relative property (T) and choose $\varepsilon > 0$. Because $\varphi_n \in c_0(\Gamma)$, we find $n \in \mathbb{N}$ and a finite subset $\mathcal{F} \subset \Gamma$ such that for all b in the unit ball of N, the $\|\cdot\|_2$-distance of b to $\mathrm{span}\{au_g \mid a \in A, g \in \mathcal{F}\}$ is smaller than ε. The only 'obvious' subalgebras of $A \rtimes \Gamma$ with such an approximation property are those that are unitarily conjugate to a subalgebra of A, i.e. $vNv^* \subset A$ for some $v \in \mathcal{U}(M)$. Popa's *intertwining-by-bimodules,* that we recall below, ensures that this feeling is indeed (almost) correct. It is even exactly correct when moreover $N \subset M$ is a Cartan subalgebra.

Now observe that $A \subset M$ has the relative property (T) and is a Cartan subalgebra. So, whenever α is an automorphism of M, the subalgebra $\alpha(A)$ still has the relative property (T) and the previous paragraph implies that, up to a unitary conjugacy, every automorphism of M globally preserves A. This means that every automorphism of M induces an automorphism of the orbit equivalence relation of $\mathrm{SL}(2,\mathbb{Z}) \curvearrowright \mathbb{T}^2$. A similar statement is true for isomorphisms $M \to pMp$ and therefore, the fundamental group of M equals the fundamental group of the equivalence relation, which is trivial by [Ga99, Ga01].

Theorem 3.6 (Popa [Po03, Po04]). *Let Γ be a property (T) group and $\Gamma \curvearrowright (X,\mu)$ any free ergodic p.m.p. action. Let Λ be any icc group and $\Lambda \curvearrowright (Y_0,\eta_0)^\Lambda$ the Bernoulli action. Put $(Y,\eta) := (Y_0,\eta_0)^\Lambda$.*

If $\mathrm{L}^\infty(X) \rtimes \Gamma \cong \mathrm{L}^\infty(Y) \rtimes \Lambda$, then the groups Γ, Λ are isomorphic and their actions are conjugate.

Put $A = \mathrm{L}^\infty(X)$ and $B_0 = \mathrm{L}^\infty(Y_0)$. Assume that $A \rtimes \Gamma = B_0^\Lambda \rtimes \Lambda$ and consider on $B_0^\Lambda \rtimes \Lambda$ the tensor length deformation θ_ρ defined after Example 3.4. Since the subalgebra $\mathrm{L}\Gamma$ has property (T), for ρ close enough to 1, we get that θ_ρ is uniformly close to the identity on the unit ball of $\mathrm{L}\Gamma$. When $b \in B_0^\Lambda$, the norm $\|\theta_\rho(b) - b\|_2$ is small when b can be written as a linear combination of 'short' elementary tensors. The only obvious 'short' subalgebras of $B_0^\Lambda \rtimes \Lambda$ are those that can be unitarily conjugated into either $\mathrm{L}\Lambda$ or B_0^J for some finite subset $J \subset \Lambda$. The abelian algebra B_0^J can never house the property (T) algebra $\mathrm{L}\Gamma$ and Popa indeed manages to prove that $\mathrm{L}\Gamma$ must be unitarily conjugate to a subalgebra of $\mathrm{L}\Lambda$.

So we may assume that $\mathrm{L}\Gamma \subset \mathrm{L}\Lambda$. In a second and analytically very delicate part, Popa basically proves the following: if a subalgebra A of $B_0^\Lambda \rtimes \Lambda$ is both abelian and normalized by many unitaries in $\mathrm{L}\Lambda$, then A must be unitarily conjugate to a subalgebra of B_0^Λ.

This brings us in the situation where A can be unitarily conjugated into B_0^Λ and $\mathrm{L}\Gamma$ can be unitarily conjugated into $\mathrm{L}\Lambda$. Popa proves that automatically both unitary conjugations can be done with the same unitary, yielding isomorphism of the groups and conjugacy of the actions.

Popa's intertwining-by-bimodules. In [Po01, Po03] Popa developed a powerful technique to approach the following question: when are two subalgebras $N, P \subset M$ unitarily conjugate? A detailed explanation and motivation for his method can be found in [Po06b, Section 5] and [BO08, Appendix F]. So, we are rather brief here. First consider the case where $M = P \rtimes \Gamma$ for some trace preserving action $\Gamma \curvearrowright P$. Every element $x \in M$ has a unique *Fourier expansion*

$$x = \sum_{g \in \Gamma} x_g u_g \quad \text{with} \quad x_g \in P$$

converging in $\|\cdot\|_2$. We call the $x_g \in P$ the Fourier coefficients of x.

Theorem 3.7 (Popa [Po01, Po03]). *Let $N \subset P \rtimes \Gamma$ be a von Neumann subalgebra. Then the following two conditions are equivalent.*

- *There exist projections $p \in P, q \in N$, a normal unital $*$-homomorphism $\varphi : qNq \to pPp$ and a non-zero partial isometry $v \in q(P \rtimes \Gamma)p$ satisfying $av = v\varphi(a)$ for all $a \in qNq$.*

- *There is no sequence of unitaries $v_n \in N$ whose Fourier coefficients converge to 0 pointwise in $\|\cdot\|_2$, i.e. $\|(v_n)_g\|_2 \to 0$ for all $g \in \Gamma$.*

When $P \subset M$ is no longer the 'core' of a crossed product, there is no notion of Fourier coefficients and their convergence to 0 has to be replaced by the condition $\|E_P(xv_ny)\|_2 \to 0$ for all $x, y \in M$, where $E_P : M \to P$ is the unique trace preserving conditional expectation. If $M = P \rtimes \Gamma$, note that $(v_n)_g = E_P(v_n u_g^*)$.

The first condition in Theorem 3.7 is of course not saying that v is a unitary satisfying $v^*Nv \subset P$. The left support projection of v lies in the relative commutant of qNq – which in concrete applications is usually known – but the right support projection of v lies in the relative commutant of $\varphi(qNq)$ which is of course a priori unknown, since we do not know φ. Several techniques based on *mixing properties* have been developed to take care of this relative commutant issue and we refer to [Po06b, Section 5] for a more detailed explanation.

4. Fundamental Groups of II_1 Factors

In order to construct II_1 factors with a prescribed fundamental group $\mathcal{F}$, you first need a cute construction of a II_1 factor M such that all $t \in \mathcal{F}$ obviously belong to $\mathcal{F}(M)$ and then you need a powerful theory to make sure that no other $t > 0$ belong to $\mathcal{F}(M)$. As an illustration, we first briefly explain two constructions that produce II_1 factors with prescribed countable fundamental group.

Connes-Størmer Bernoulli actions [Po03]. Let (X_0, μ_0) be an atomic probability space and Γ a countable group. Put $(X, \mu) = (X_0, \mu_0)^\Gamma$ and define on (X, μ) the following II_1 equivalence relation: $x \sim y$ if and only if there exists a $g \in \Gamma$ and a finite subset $J \subset \Gamma$ such that $x_{gh} = y_h$ for all $h \in \Gamma - J$ and such that $\prod_{h \in J} \mu_0(x_{gh}) = \prod_{h \in J} \mu_0(y_h)$. Whenever $a \in X_0$, define the subset $Y_a \subset X$ as $Y_a := \{x \in X \mid x_e = a\}$. Given $a, b \in X_0$, the map $Y_a \to Y_b$ changing x_e from a to b and leaving the other x_g untouched is an isomorphism of the restricted equivalence relations. Hence, $\mu_0(a)/\mu_0(b)$ belongs to $\mathcal{F}(\mathcal{R})$ for all $a, b \in X_0$.

Denote by $M = \mathrm{L}\mathcal{R}$ the II_1 factor associated with $\mathcal{R}$ and denote by $\mathcal{F}$ the subgroup of $\mathbb{R}_+^*$ generated by all the ratios $\mu_0(a)/\mu_0(b)$. We have seen that $\mathcal{F} \subset \mathcal{F}(M)$. In [Po03] Popa shows that taking $\Gamma = \mathrm{SL}(2, \mathbb{Z}) \ltimes \mathbb{Z}^2$, one has the equality $\mathcal{F} = \mathcal{F}(M)$, with one of the ingredients of the proof being the triviality of the fundamental group of $\mathrm{L}\Gamma$.

Free products of amplifications [IPP05]. Recall from Section 2 the notation M^t for the amplification of a II_1 factor M. In [DR98] Dykema and Rădulescu established the following remarkable formula for an infinite free product of II_1 factors M_n:

$$\Bigl(\mathop{*}_{n \in \mathbb{N}} M_n\Bigr)^t \cong \mathop{*}_{n \in \mathbb{N}} M_n^t \; .$$

So, whenever $\mathcal{F} \subset \mathbb{R}_+^*$ is a countable subgroup different from $\{1\}$ and whenever M is a II_1 factor, we put $P = *_{t \in \mathcal{F}} M^t$ and conclude that both $\mathcal{F}$ and $\mathcal{F}(M)$ are subgroups of $\mathcal{F}(P)$. As a consequence of their Bass-Serre theory for free products of II_1 factors, Ioana, Peterson and Popa [IPP05] manage to prove that $\mathcal{F} = \mathcal{F}(P)$ when you take $M = \mathrm{L}(\mathrm{SL}(2, \mathbb{Z}) \ltimes \mathbb{Z}^2)$. A more elementary variant of the previous construction was proposed in [Ho07].

Both constructions presented so far can be used to produce II_1 factors with arbitrary fundamental group $\mathcal{F}$. However, for uncountable subgroups $\mathcal{F} \subset \mathbb{R}_+^*$ the resulting II_1 factors do not have separable predual.

II_1 factors with uncountable fundamental group [PV08a, PV08c]. Again we start with a construction of a II_1 factor sowing a number of elements into the fundamental group. Let $\Gamma \curvearrowright (Z, \gamma)$ be a free ergodic action preserving the *infinite* non-atomic measure γ. Put $N := \mathrm{L}^\infty(Z) \rtimes \Gamma$ and note that N is a II_∞ factor. Whenever Δ is a non-singular[10] automorphism of (Z, γ) satisfying $\Delta(g \cdot z) = g \cdot \Delta(z)$ a.e., the Radon-Nikodym derivative between $\gamma \circ \Delta^{-1}$ and γ is Γ-invariant and hence constant a.e. So, Δ scales the infinite measure γ by a factor that we denote by $\mathrm{mod}\, \Delta$. Denote the group of all these automorphisms Δ as $\mathrm{Centr}_{\mathrm{Aut}\, Z}(\Gamma)$. Every $\Delta \in \mathrm{Centr}_{\mathrm{Aut}\, Z}(\Gamma)$ gives rise to an automorphism Δ_* of N scaling the trace with the same factor $\mathrm{mod}\, \Delta$. Whenever $p \in N$ is a

[10] This means that Δ preserves sets of measure zero.

projection of finite trace, put $M = pNp$ and note that M is a II_1 factor. By construction we have

$$\mathrm{mod}\,\mathrm{Centr}_{\mathrm{Aut}\,Z}(\Gamma) \subset \mathcal{F}(pNp)\ . \tag{1}$$

In the following paragraphs we explain the three main aspects of [PV08a, PV08c]: how to get an equality in (1), how to make sure that pNp is itself a group measure space II_1 factor and finally, how wild $\mathrm{mod}\,\mathrm{Centr}_{\mathrm{Aut}\,Z}(\Gamma)$ can be. At the end this will give a feeling for the validity of the following theorem.

Theorem 4.1 (Popa, Vaes [PV08a, PV08c]). *Let Γ_0 be a non-trivial group and Σ an infinite amenable group. Put $\Gamma = \Gamma_0^{*\infty} * \Sigma$. There exists a free ergodic p.m.p. action $\Gamma \curvearrowright (X,\mu)$ such that $\mathrm{L}^\infty(X) \rtimes \Gamma$ has a fundamental group of arbitrary prescribed Hausdorff dimension between 0 and 1.*

Actually, $\mathcal{F}(\mathrm{L}^\infty(X) \rtimes \Gamma)$ can be any group in the family

$$\mathcal{S}_{\mathrm{centr}} := \{\mathcal{F} \subset \mathbb{R}_+^* \mid \textit{there exists an amenable } \Lambda \textit{ and a free ergodic } \Lambda \curvearrowright (Y,\eta) \textit{ preserving } \eta \textit{ such that } \mathcal{F} = \mathrm{mod}\,\mathrm{Centr}_{\mathrm{Aut}\,Y}(\Lambda)\ \}$$

How to get equality in (1)**.** Take Γ of the form $\Gamma_1 * \Lambda$ and assume that $\Gamma \curvearrowright (X,\mu)$ is a free p.m.p. action with Γ_1 acting ergodically. Let $\Lambda \curvearrowright (Y,\eta)$ be a free ergodic action preserving the infinite non-atomic measure η. Put $(Z,\gamma) = (X \times Y, \mu \times \eta)$ and consider the action $\Gamma \curvearrowright Z$ given by $g \cdot (x,y) = (g \cdot x, y)$ if $g \in \Gamma_1$ and $h \cdot (x,y) = (h \cdot x, h \cdot y)$ if $h \in \Lambda$. It is easy to see that $\mathrm{mod}\,\mathrm{Centr}_{\mathrm{Aut}\,Z}(\Gamma)$ equals $\mathrm{mod}\,\mathrm{Centr}_{\mathrm{Aut}\,Y}(\Lambda)$. We now make the following assumptions on the action $\Gamma_1 * \Lambda \curvearrowright (X,\mu)$.

1. The action $\Gamma_1 \curvearrowright (X,\mu)$ is rigid, i.e. $\mathrm{L}^\infty(X) \subset \mathrm{L}^\infty(X) \rtimes \Gamma_1$ has the relative property (T) in the sense of Definition 2.1.
2. The group Λ is amenable.
3. If $\phi : X_0 \to X_1$ is a non-singular isomorphism between the non-negligible subsets $X_0, X_1 \subset X$ satisfying $\phi(X_0 \cap \Gamma_1 \cdot x) \subset \Gamma \cdot \varphi(x)$ for a.e. $x \in X_0$, then $\phi(x) \in \Gamma \cdot x$ for a.e. $x \in X_0$.

For the following heuristic reasons, these assumptions imply that the inclusion in (1) actually is an equality. Consider the II_∞ factor $N = \mathrm{L}^\infty(X \times Y) \rtimes (\Gamma_1 * \Lambda)$ and let α be an automorphism of N. The Bass-Serre theory of [IPP05] and the relative property (T) of $\mathrm{L}^\infty(X)$ imply that after a unitary conjugacy, α globally preserves $\mathrm{L}^\infty(X \times Y)$. In a next step, the relative property (T) of $\mathrm{L}^\infty(X)$ together with the amenability of Λ roughly implies that α globally preserves $\mathrm{L}^\infty(X) \otimes 1$. Then the third assumption above implies that we may assume that $\alpha(a \otimes 1) = a \otimes 1$ for all $a \in \mathrm{L}^\infty(X)$ and hence that α is induced by an element of $\mathrm{Centr}_{\mathrm{Aut}\,Y}(\Lambda)$.

It is highly non-trivial to find group actions $\Gamma_1 * \Lambda \curvearrowright (X,\mu)$ satisfying the conditions 1, 2 and 3 above. Actually, in [PV08a, PV08c] we use a Baire category

argument to prove their existence whenever Γ_1 is an infinite free product $\Gamma_1 = \Gamma_0^{*\infty}$ with Γ_0 being an arbitrary non-trivial group. Replacing Γ_0 by $\Gamma_0 * \Gamma_0$, we may assume that Γ_0 is infinite. The proof roughly goes as follows: start with arbitrary free ergodic p.m.p. actions $\Gamma_0 \curvearrowright (X,\mu)$ and $\Lambda \curvearrowright (X,\mu)$ and view Γ, Λ as subgroups of $\mathrm{Aut}(X,\mu)$. By [Ga08], there exists an automorphism $\beta_1 \in \mathrm{Aut}(X,\mu)$ such that the subgroups Γ_0 and $\beta_1\Gamma_0\beta_1^{-1}$ of $\mathrm{Aut}(X,\mu)$ generate a *free and rigid* action of $\Gamma_0^{*2} := \Gamma_0 * \Gamma_0$ on (X,μ). By [IPP05, To05] there exists $\psi \in \mathrm{Aut}(X,\mu)$ such that together with $\psi\Lambda\psi^{-1}$, we obtain a free action of $\Gamma_0^{*2} * \Lambda$ on (X,μ). We now start adding copies of Γ_0 acting as $\beta_n\Gamma_0\beta_n^{-1} \subset \mathrm{Aut}(X,\mu)$ for well chosen $\beta_n \in \mathrm{Aut}(X,\mu)$. At stage n, given the free action $\Gamma_0^{*n} * \Lambda \curvearrowright X$, the rigidity of the action $\Gamma_0^{*2} \curvearrowright (X,\mu)$ implies that there are essentially only countably many partial isomorphisms ϕ that map Γ_0^{*2}-orbits into $\Gamma_0^{*n} * \Lambda$-orbits. By [IPP05, To05] there exists $\beta_{n+1} \in \mathrm{Aut}(X,\mu)$ such that $\beta_{n+1}\Gamma_0\beta_{n+1}^{-1}$ is free w.r.t. these countably many partial isomorphisms. Adding this new Γ_0-action, we obtain a free action of $\Gamma_0^{*(n+1)} * \Lambda$. Continuing by induction, we get the free action $\Gamma_0^{*\infty} * \Lambda \curvearrowright (X,\mu)$. We finally prove that there exists an infinite subset $E \subset \mathbb{N}$ containing $\{0,1\}$ such that $(*_{n\in E}\Gamma_0) * \Lambda \curvearrowright (X,\mu)$ satisfies condition 3. Since $\{0,1\} \subset E$ and $\Gamma_0^{*2} \curvearrowright (X,\mu)$ is rigid, condition 1 is satisfied as well.

Is *pNp* itself a group measure space factor. Take $\Gamma_1 * \Lambda \curvearrowright X \times Y$ as above, with Λ being infinite amenable. Let Σ be any infinite amenable group. By [PV08c, Lemma 3.6], the restriction of the orbit equivalence relation $\mathcal{R}(\Gamma_1 * \Lambda \curvearrowright X \times Y)$ to a set of finite measure is implemented by a free action of $\Gamma_1^{*\infty} * \Sigma$. So, pNp is a group measure space factor.

How wild can $\mathbf{\mathrm{mod}\,\mathrm{Centr}_{\mathrm{Aut}\,Y}(\Lambda)}$ be. In [MNP68] the notion of an *ergodic measure* ν on the real line is introduced: it is a σ-finite measure on the Borel sets of $\mathbb{R}$ such that for every $x \in \mathbb{R}$, the translation ν_x of ν by x is either singular w.r.t. ν or equal to ν, and such that denoting

$$H_\nu := \{x \in \mathbb{R} \mid \nu_x = \nu\}$$

the action of H_ν on $(\mathbb{R},\nu)$ by translation is ergodic. More precisely, if $F : \mathbb{R} \to \mathbb{R}$ is a Borel function and if for every $x \in H_\nu$ we have $F(x+y) = F(y)$ for ν-a.e. y, then F is constant ν-a.e. The easiest examples of ergodic measures are the Lebesgue measure and the counting measure on a countable subgroup of $\mathbb{R}$.

In [Aa86] it is shown that all subgroups of $\mathbb{R}_+^*$ of the form $\exp(H_\nu)$ arise as $\mathrm{mod}\,\mathrm{Centr}_{\mathrm{Aut}\,Y}(\mathbb{Z})$. Allowing more general amenable groups Λ instead of $\mathbb{Z}$, this can be easily seen as follows. Assume that $H_\nu \neq \{0\}$. Viewing H_ν as a closed subgroup of $\mathrm{Aut}(\mathrm{L}^\infty(\mathbb{R},\nu))$ one turns H_ν into a Polish group. Take a countable dense subgroup $Q \subset H_\nu$ and define the additive subgroup $R \subset \mathbb{R}$ as $R = \mathbb{Z}[\exp(Q)]$. Then Q acts on R through multiplication by $\exp(q)$ and we put $\Lambda := R \rtimes Q$. Define the measure $\widetilde{\nu}$ on $\mathbb{R}$ such that $d\widetilde{\nu}(x) = \exp(-x)d\nu(x)$ and denote by λ the Lebesgue measure on $\mathbb{R}$. Put $(Y,\eta) := (\mathbb{R}^2, \lambda \times \widetilde{\nu})$. The action

$\Lambda \curvearrowright Y$ given by $(r,q)\cdot(x,y) = (r+\exp(q)x, q+y)$ is measure preserving and ergodic. One checks easily that $\operatorname{mod}\operatorname{Centr}_{\operatorname{Aut} Y}(\Lambda) = \exp(H_\nu)$.

It is also shown in [Aa86] (cf. [PV08a, page 389]) that H_ν can be uncountable without being $\mathbb{R}_+^*$ and that actually H_ν can have any Hausdorff dimension between 0 and 1.

Problem I. Give an intrinsic description of the subgroups of $\mathbb{R}_+^*$ that are of the form $\mathcal{F}(M)$ where M is a II_1 factor with separable predual.

The only known a priori restriction on $\mathcal{F}(M)$ is given by [PV08a, Proposition 2.1]: $\mathcal{F}(M)$ is a Borel subset of $\mathbb{R}_+^*$ and carries a unique Polish group topology whose Borel sets are precisely the ones inherited from $\mathbb{R}_+^*$. It is however hard to believe that all 'Polishable' Borel subgroups of $\mathbb{R}_+^*$ can arise as the fundamental group of a II_1 factor with separable predual.

Property (T) and fundamental groups. Roughly speaking, the presence of property (T) forces fundamental groups to be countable. When Γ is an icc property (T) group, Connes [Co80] proved that $\mathcal{F}(\mathrm{L}\Gamma)$ is countable and the same method [GG88] yields the countability of $\mathcal{F}(\mathrm{L}^\infty(X)\rtimes\Gamma)$ and $\mathcal{F}(\mathcal{R}(\Gamma\curvearrowright X))$, for all free ergodic p.m.p. actions $\Gamma\curvearrowright(X,\mu)$. For non-icc property (T) groups, Ioana [Io08] could still prove the countability of $\mathcal{F}(\mathcal{R}(\Gamma\curvearrowright X))$, but in [PV08c] we proved that if Γ is a property (T) group whose virtual center[11] is not virtually abelian[12], then Γ admits a free ergodic p.m.p. action such that $\mathrm{L}^\infty(X)\rtimes\Gamma$ is McDuff and in particular, has fundamental group $\mathbb{R}_+^*$. We were informed by Ershov that such groups exist as quotients of Golod-Shafarevich groups with property (T).

Zimmer [Zi80] introduced a notion of property (T) for II_1 equivalence relations which is such that for free ergodic p.m.p. actions, the orbit equivalence relation $\mathcal{R}(\Gamma\curvearrowright X)$ has property (T) if and only if the group Γ has property (T). Using techniques from [AD04] we proved [PV08b] that for $n\geq 4$, the restriction of the II_∞ relation $\mathcal{R}(\mathrm{SL}(n,\mathbb{Z})\curvearrowright\mathbb{R}^n)$ to a subset of finite Lebesgue measure, is a II_1 equivalence relation with property (T) in the sense of Zimmer and with fundamental group $\mathbb{R}_+^*$. In particular, this II_1 equivalence relation cannot be implemented by a free action of a group.

5. (Non-)uniqueness of Cartan Subalgebras

Non-uniqueness of Cartan subalgebras. Connes and Jones [CJ81] have given the first examples of II_1 factors M having more than one Cartan subalgebra up to conjugacy by an automorphism of M. Their construction goes as

[11] The virtual center is the set of elements with finite conjugacy class. It is a normal subgroup.

[12] A group is virtually abelian if it has an abelian subgroup of finite index.

follows. Take a finite non-abelian group Σ_0, build the countable group $\Sigma = \Sigma_0^{(\mathbb{N})}$ and the compact group $K = \Sigma_0^{\mathbb{N}}$ that we equip with its Haar measure. Consider the action $\Sigma \curvearrowright K$ by translation. Finally, let Γ be any non-amenable group. Put $X := K^\Gamma$ and consider the diagonal action $\Sigma \curvearrowright X$ which commutes with the Bernoulli action $\Gamma \curvearrowright X$. We obtain a free ergodic p.m.p. action $\Gamma \times \Sigma \curvearrowright X$ and hence, the Cartan subalgebra $A = \mathrm{L}^\infty(X)$ of $M = \mathrm{L}^\infty(X) \rtimes (\Gamma \times \Sigma)$. Taking non-commuting elements $g, h \in \Sigma_0$ and defining $g_n, h_n \in \Sigma$ as being g, h in position n, one obtains two non-commuting central sequences in M. Hence, M is a *McDuff factor* [McD70], which means that $M \cong M \overline{\otimes} R$, where R denotes the hyperfinite II_1 factor. Choosing any Cartan subalgebra $B \subset R$, one transports back the Cartan subalgebra $A \overline{\otimes} B \subset M \overline{\otimes} R$ to a Cartan subalgebra of M whose associated equivalence relation is not strongly ergodic. The initial equivalence relation given by $A \subset M$ is strongly ergodic, so that both Cartan subalgebras are non-conjugate by an automorphism of M.

Ozawa and Popa [OP07] provided examples of II_1 factors where one can explicitly see two Cartan subalgebras. We first explain the general procedure and provide a concrete example below. Assume that Γ is a countable group and $\Sigma \lhd \Gamma$ an infinite abelian normal subgroup. Assume that $\Sigma \lhd \Gamma$ has the following relative icc property: for every $g \in \Gamma - \Sigma$, the set $\{sgs^{-1} \mid s \in \Sigma\}$ is infinite. Let $\Sigma \hookrightarrow K$ be a dense embedding of Σ into a compact abelian group K. Equip K with its Haar measure and define the action $\Sigma \curvearrowright K$ by translation. Assume that we are given an extension of this action to an essentially free p.m.p. action $\Gamma \curvearrowright K$. Then, $\mathrm{L}^\infty(K)$ and $\mathrm{L}\Sigma$ are non-unitarily conjugate Cartan subalgebras of $\mathrm{L}^\infty(K) \rtimes \Gamma$.

An interesting concrete example is given by $\Gamma = \mathbb{Z}^n \rtimes \mathrm{SL}(n, \mathbb{Z})$ and its natural affine action on $\mathbb{Z}_p^n$. We get

$$\mathrm{L}^\infty(\mathbb{Z}_p^n) \rtimes (\mathbb{Z}^n \rtimes \mathrm{SL}(n, \mathbb{Z})) = \mathrm{L}^\infty(\mathbb{T}^n) \rtimes (\widehat{\mathbb{Z}_p^n} \rtimes \mathrm{SL}(n, \mathbb{Z})) \; .$$

If $n = 2$, the group $\widehat{\mathbb{Z}_p^n} \rtimes \mathrm{SL}(n, \mathbb{Z})$ has the Haagerup property while $\mathbb{Z}^n \rtimes \mathrm{SL}(n, \mathbb{Z})$ does not. If $n \geq 3$, the group $\mathbb{Z}^n \rtimes \mathrm{SL}(n, \mathbb{Z})$ has property (T) while $\widehat{\mathbb{Z}_p^n} \rtimes \mathrm{SL}(n, \mathbb{Z})$ does not. So, neither property (T) nor the Haagerup property are stable under W*-equivalence. Since they are stable under orbit equivalence, it follows that the Cartan subalgebras $\mathrm{L}^\infty(\mathbb{Z}_p^n)$ and $\mathrm{L}^\infty(\mathbb{T}^n)$ are non-conjugate by an automorphism of the ambient II_1 factor.

Uniqueness of Cartan subalgebras. We mentioned in Section 1 that Ozawa and Popa [OP07, OP08] established the first – and up to now, only – uniqueness results for Cartan subalgebras up to unitary conjugacy. Recall that a *profinite* p.m.p. action $\Gamma \curvearrowright (X, \mu)$ is by definition the inverse limit $\varprojlim (X_n, \mu_n)$ of a directed system of actions $\Gamma \curvearrowright (X_n, \mu_n)$ on *finite* probability spaces.

Theorem 5.1 (Ozawa, Popa [OP07]). *Let $n \geq 2$ and let $\mathbb{F}_n \curvearrowright (X, \mu)$ be an ergodic profinite p.m.p. action. Put $A = \mathrm{L}^\infty(X)$ and $M = A \rtimes \mathbb{F}_n$.*

It the action is free, $A \subset M$ is the unique Cartan subalgebra up to unitary conjugacy. If the action is not free, M has no Cartan subalgebras.

The proof of Theorem 5.1 consists of two parts. The group $\mathbb{F}_n$ has the *complete metric approximation property* (CMAP) [Ha78]. This means that there exists a sequence $\varphi_k : \Gamma \to \mathbb{C}$ of finitely supported functions converging to 1 pointwise and such that the maps $\theta_k : u_g \mapsto \varphi_k(g)u_g$ are *completely bounded* with $\limsup \|\theta_k\|_{\mathrm{cb}} = 1$. Then, θ_k automatically extends to an ultraweakly continuous map $\mathrm{L}\mathbb{F}_n \to \mathrm{L}\mathbb{F}_n$ without increasing $\|\theta_k\|_{\mathrm{cb}}$. Since $\mathbb{F}_n \curvearrowright (X, \mu)$ is profinite, it follows that M has CMAP as a II_1 factor: there exists a sequence of finite rank, ultraweakly continuous, completely bounded maps $\theta_k : M \to M$ converging pointwise in $\|\cdot\|_2$ to the identity and satisfying $\limsup \|\theta_k\|_{\mathrm{cb}} = 1$. Ozawa and Popa prove the following general statement: let M be a II_1 factor with CMAP and $P \subset M$ a diffuse amenable (in particular, abelian) subalgebra; denote by $\mathcal{G}$ the group of unitaries in M normalizing P. Then, the action of $\mathcal{G}$ on P by automorphisms $\mathrm{Ad}\, u$ is necessarily *weakly compact*. We do not explain this notion here but just note that the slightly stronger notion of *compactness* means that the closure of $\mathrm{Ad}\,\mathcal{G}$ inside the Polish group $\mathrm{Aut}\, P$ is compact.

So far, the reasoning in the previous paragraph works for any profinite action of a CMAP group. In the second part of the proof, Ozawa and Popa prove the following. Let $\mathbb{F}_n \curvearrowright (X, \mu)$ be any ergodic p.m.p. action, $P \subset \mathrm{L}^\infty(X) \rtimes \mathbb{F}_n$ a subalgebra and $\mathcal{G}$ a group of unitaries normalizing P such that the action $\mathrm{Ad}\,\mathcal{G}$ on P is weakly compact. Then either $\mathcal{G}$ generates an amenable von Neumann algebra or there almost exists a unitary conjugacy of P into $\mathrm{L}^\infty(X)$ (in the sense of Theorem 3.7). The two parts together yield Theorem 5.1.

Problem II. Let $n \geq 2$ and let $\mathbb{F}_n \curvearrowright (X, \mu)$ be *any* free ergodic p.m.p. action. Does $\mathrm{L}^\infty(X) \rtimes \mathbb{F}_n$ always have a unique Cartan subalgebra up to unitary conjugacy?

Let $n \geq 2$ and let M be an arbitrary II_1 factor. Is it always true that $M \overline{\otimes} \mathrm{L}\mathbb{F}_n$ has no Cartan subalgebra? By [OP07], this is indeed the case if M has the complete metric approximation property.

Another breakthrough unique Cartan decomposition theorem was obtained in [PV09]. On the one hand it is weaker than Theorem 5.1 since we are only able to deal with group measure space Cartan subalgebras, but on the other hand it is stronger since we consider arbitrary group actions.

Theorem 5.2 (Popa, Vaes [PV09]). *Let $\Gamma = \Gamma_1 * \Gamma_2$ be a free product where Γ_1 admits a non-amenable subgroup with the relative property (T) and where Γ_2 is any non-trivial group. Let $\Gamma \curvearrowright (X, \mu)$ be any ergodic p.m.p. action and put $M = \mathrm{L}^\infty(X) \rtimes \Gamma$.*

If the action is free, $\mathrm{L}^\infty(X)$ is, up to unitary conjugacy, the unique group measure space Cartan subalgebra. If the action is not free, M has no group measure space Cartan subalgebra.

Write $A = \mathrm{L}^\infty(X)$. So, $M = A \rtimes \Gamma$ and we denote by $x = \sum_{g\in\Gamma} x_g u_g$ the unique Fourier expansion of $x \in M$, with $x_g \in A$ for all $g \in \Gamma$.

Assume that $M = B \rtimes \Lambda$ is any other group measure decomposition. Denote by $(v_s)_{s\in\Lambda}$ the canonical unitary elements that correspond to this decomposition. The first step of the proof of Theorem 5.2 consists in transferring some of the rigidity of Γ to Λ. Assume that φ_n is a deformation of the identity of M. If instead of Γ, our unknown group Λ would have a non-amenable subgroup Λ_0 with the relative property (T), this would imply that on the unitary elements $v_s, s \in \Lambda_0$, the deformation φ_n converges uniformly to the identity: for n large enough and all $s \in \Lambda_0$, we get that $\|\varphi_n(v_s) - v_s\|_2$ is small. Moreover, the abelian algebra A cannot contain a copy of the non-amenable algebra $\mathrm{L}\Lambda_0$ so that Theorem 3.7 provides a sequence s_k in Λ_0 such that the Fourier coefficients of v_{s_k} tend to zero pointwise in $\|\cdot\|_2$: for all $g \in \Gamma$, we have $\|(v_{s_k})_g\|_2 \to 0$ as $k \to \infty$.

Obviously, we cannot prove that Λ automatically has a non-amenable subgroup with the relative property (T). Nevertheless, we have the following transfer of rigidity result.

Lemma 5.3. *Let $A \rtimes \Gamma = M = B \rtimes \Lambda$ be two crossed product decompositions of the same II_1 factor with A and B being amenable. Assume that Γ admits a non-amenable subgroup with the relative property (T) and let φ_n be a deformation of the identity of M. For every $\varepsilon > 0$ there exists an $n \in \mathbb{N}$ and a sequence of group elements $s_k \in \Lambda$ such that*

- *$\|\varphi_n(v_{s_k}) - v_{s_k}\|_2 < \varepsilon$ for all k,*
- *the Fourier coefficients of v_{s_k} tend to zero pointwise: for all $g \in \Gamma$, we have $\|(v_{s_k})_g\|_2 \to 0$ as $k \to \infty$.*

Lemma 5.3 is proven by playing the following positive-definite ping-pong: the formula $\psi_n(s) = \tau(v_s^* \varphi_n(v_s))$ defines a sequence of positive definite functions on Λ, which in turn define completely positive maps $\theta_n : B \rtimes \Lambda \to B \rtimes \Lambda : \theta_n(bv_s) = \psi_n(s) b v_s$, which in their turn define positive definite functions $\gamma_n : \Gamma \to \mathbb{C} : \gamma_n(g) = \tau(u_g^* \theta_n(u_g))$. By construction, $\gamma_n \to 1$ pointwise and hence uniformly on the subgroup of Γ with the relative property (T). From this, one can deduce the conclusion of Lemma 5.3.

The starting point to prove Theorem 5.2 is an application of Lemma 5.3 to the word length deformation φ_ρ on $M = A \rtimes (\Gamma_1 * \Gamma_2)$ given in Example 3.2. The fact that $\|\varphi_\rho(x) - x\|_2$ is small means that x lies close to a linear combination of au_g, $a \in A$ and $g \in \Gamma$ with $|g|$ not too large. We refer to such elements $x \in M$ as being 'short'. Lemma 5.3 provides a sequence of short unitaries v_{s_k}. Since B is abelian and normalized by the short unitaries v_{s_k}, a combinatorial argument implies that the elements of B are themselves 'uniformly short'. But Ioana, Peterson and Popa, in their study of rigid subalgebras of amalgamated free products [IPP05], proved that this implies that B can be unitarily conjugated into one of the 'obvious' short subalgebras of $A \rtimes (\Gamma_1 * \Gamma_2)$, namely $A \rtimes \Gamma_1$ or

$A \rtimes \Gamma_2$. It follows that B can actually be conjugated into A since otherwise the normalizer of B, i.e. the whole of M, would get conjugated in $A \rtimes \Gamma_i$ as well.

Problem III. Does the transfer of rigidity lemma hold for arbitrary Cartan subalgebras? More precisely, let $B \subset M$ be a Cartan subalgebra. Does there exist an n and a sequence of unitaries $v_k \in M$, normalizing B and satisfying the same two properties as the unitaries v_{s_k} in the formulation of Lemma 5.3?

In the affirmative case, one can replace 'group measure space Cartan' by 'Cartan' throughout the formulation of Theorem 5.2.

6. Superrigidity for Group Measure Space Factors

As explained in paragraph 1.6, W*-superrigidity of an action $\Gamma \curvearrowright (X, \mu)$ arises as the 'sum' of OE superrigidity and the uniqueness of the group measure space Cartan subalgebra in $\mathrm{L}^\infty(X) \rtimes \Gamma$, up to unitary conjugacy. This makes W*-superrigidity theorems extremely hard to obtain.

Unfortunately, none of the profinite actions covered by the uniqueness of Cartan theorems in [OP07, OP08] is known to be (virtually) OE superrigid. Also actions $\Gamma_1 * \Gamma_2 \curvearrowright (X, \mu)$ are not OE superrigid so that we cannot directly apply Theorem 5.2. But Theorem 5.2 can be generalized so that it covers arbitrary actions of certain *amalgamated free products* $\Gamma = \Gamma_1 *_\Sigma \Gamma_2$ over a common amenable subgroup Σ which is sufficiently non-normal inside Γ (see [PV09]). In combination with Popa's OE superrigidity for Bernoulli actions [Po05] or Kida's OE superrigidity [Ki09], we obtained the following examples of W*-superrigid actions. A general statement and more examples, including Gaussian actions and certain co-induced actions, can be found in [PV09].

Theorem 6.1 (Popa, Vaes [PV09]). *Let $n \geq 3$ and denote by $T_n < \mathrm{PSL}(n, \mathbb{Z})$ the subgroup of upper triangular matrices.*

Let $\Lambda \neq \{e\}$ be an arbitrary group and $\Sigma < T_n$ an infinite subgroup. Put $\Gamma = \mathrm{PSL}(n, \mathbb{Z}) *_\Sigma (\Sigma \times \Lambda)$. Then, the Bernoulli action $\Gamma \curvearrowright (X_0, \mu_0)^\Gamma$ is W-superrigid. More generally, whenever $\Gamma \curvearrowright I$ and $\Sigma \cdot i$ is infinite for all $i \in I$, the generalized Bernoulli action $\Gamma \curvearrowright (X_0, \mu_0)^I$ is W*-superrigid.*

Any free mixing p.m.p. action of $\mathrm{PSL}(n, \mathbb{Z}) *_{T_n} \mathrm{PSL}(n, \mathbb{Z})$ is W-superrigid.*

The following beautiful result was obtained very recently by Ioana [Io10] and should be compared with Theorem 3.6.

Theorem 6.2 (Ioana [Io10]). *Let Γ be an icc group that admits an infinite normal subgroup with the relative property (T). Then, the Bernoulli action $\Gamma \curvearrowright (X_0, \mu_0)^\Gamma$ is W*-superrigid.*

Again, because of Popa's OE superrigidity for Bernoulli actions [Po05], the issue is to prove that M has a unique group measure space Cartan subalgebra

up to unitary conjugacy. Write $A = \mathrm{L}^\infty\big(X_0^\Gamma\big)$ and $M = A \rtimes \Gamma$. Assume that $M = B \rtimes \Lambda$ is another group measure space decomposition. The first step of the proof of Theorem 6.2 is similar to the transfer of rigidity ping-pong technique that I explained after Lemma 5.3. Denote by $(v_s)_{s\in\Lambda}$ the canonical group of unitaries in $B \rtimes \Lambda$. So, Ioana considers the $*$-homomorphism $\Delta : B \rtimes \Lambda \to (B \rtimes \Lambda)\overline{\otimes}\mathrm{L}\Lambda$ given by $\Delta(bv_s) = bv_s \otimes v_s$ for all $b \in B$, $s \in \Lambda$. Since $\mathrm{L}\Lambda \subset M$, we rather view Δ as an embedding of M into $M\overline{\otimes}M$.

But M arises from a Bernoulli action of a (relative) property (T) group. As a corollary of Theorem 3.6, every automorphism of M is of a special form, i.e. the composition of an inner automorphism and automorphisms induced by a character $\omega : \Gamma \to S^1$ and by a self-conjugacy of the action $\Gamma \curvearrowright A$. In an amazing technical tour de force Ioana manages to generalize Popa's methods from automorphisms to embeddings and to give an almost complete picture of all possible embeddings of M into $M\overline{\otimes}M$, when M arises from a Bernoulli action of a property (T) group. This picture is sufficiently precise so that applied to the embedding Δ constructed in the previous paragraph, one can conclude that A and B are unitarily conjugate.

References

[Aa86] J. Aaronson, The intrinsic normalising constants of transformations preserving infinite measures. *J. Analyse Math.* **49** (1987), 239–270.

[AD04] C. Anantharaman-Delaroche, Cohomology of property T groupoids and applications. *Ergodic Theory Dynam. Systems* **25** (2005), 977–1013.

[BO08] N. Brown and N. Ozawa, C^*-algebras and finite-dimensional approximations. *Graduate Studies in Mathematics* **88**, American Mathematical Society, Providence, 2008.

[CH08] I. Chifan and C. Houdayer, Bass-Serre rigidity results in von Neumann algebras. To appear in *Duke Math J.* `arXiv:0805.1566`

[Co72] A. Connes, Une classification des facteurs de type III. *Ann. Sci. École Norm. Sup.* **6** (1973), 133–252.

[Co75a] A. Connes, Sur la classification des facteurs de type II. *C. R. Acad. Sci. Paris Sér. A-B* **281** (1975), A13–A15.

[Co75b] A. Connes, Classification of injective factors. *Ann. of Math. (2)* **104** (1976), 73–115.

[Co80] A. Connes, A factor of type II_1 with countable fundamental group. *J. Operator Theory* **4** (1980), 151–153.

[Co94] A. Connes, Noncommutative geometry. Academic Press, San Diego, 1994.

[CJ81] A. Connes and V.F.R. Jones, A II_1 factor with two non-conjugate Cartan subalgebras, *Bull. Amer. Math. Soc.* **6** (1982), 211–212.

[CJ83] A. Connes and V.F.R. Jones, Property (T) for von Neumann algebras, *Bull. London Math. Soc.* **17** (1985), 57–62.

[CT76] A. Connes and M. Takesaki, The flow of weights on factors of type III. *Tôhoku Math. J.* **29** (1977), 473–575.

[CH88] M. Cowling and U. Haagerup, Completely bounded multipliers of the Fourier algebra of a simple Lie group of real rank one. *Invent. Math.* **96** (1989), 507–549.

[DV10] S. Deprez and S. Vaes, A classification of all finite index subfactors for a class of group measure space II_1 factors. *Preprint.* `arXiv:1002.2129`

[Dy58] H.A. Dye On groups of measure preserving transformation, I. *Amer. J. Math.* **81** (1959), 119–159.

[DR98] K. Dykema and F. Rădulescu, Compressions of free products of von Neumann algebras. *Math. Ann.* **316** (2000), 61–82.

[FM75] J. Feldman and C.C. Moore, Ergodic Equivalence Relations, Cohomology, and Von Neumann Algebras, I, II. *Trans. Amer. Math. Soc.* **234** (1977), 289–324, 325–359.

[FV07] S. Falguières and S. Vaes, Every compact group arises as the outer automorphism group of a II_1 factor. *J. Funct. Anal.* **254** (2008), 2317–2328.

[FV08] S. Falguières and S. Vaes, The representation category of any compact group is the bimodule category of a II_1 factor. *J. Reine Angew. Math.,* to appear. `arXiv:0811.1764`

[Fu98a] A. Furman, Gromov's measure equivalence and rigidity of higher rank lattices. *Ann. of Math.* **150** (1999), 1059–1081.

[Fu98b] A. Furman, Orbit equivalence rigidity. *Ann. of Math.* **150** (1999), 1083–1108.

[Fu09] A. Furman, A survey of measured group theory. In *Geometry, rigidity and group actions,* Eds. B. Farb and D. Fisher. The University of Chicago Press, to appear. `arXiv:0901.0678`

[Ga99] D. Gaboriau, Coût des relations d'équivalence et des groupes. *Invent. Math.* **139** (2000), 41–98.

[Ga01] D. Gaboriau, Invariants ℓ^2 de relations d'équivalence et de groupes. *Publ. Math. Inst. Hautes Études Sci.* **95** (2002), 93–150.

[Ga08] D. Gaboriau, Relative property (T) actions and trivial outer automorphism groups. *J. Funct. Anal.,* to appear. `arXiv:0804.0358`

[Ga10] D. Gaboriau, Orbit equivalence and measured group theory. In *Proceedings of the International Congress of Mathematicians (Hyderabad, 2010).*

[Ge96] L. Ge, Applications of free entropy to finite von Neumann algebras, II. *Ann. of Math.* **147** (1998), 143–157.

[GG88] S.L. Gefter and V.Ya. Golodets, Fundamental groups for ergodic actions and actions with unit fundamental groups. *Publ. Res. Inst. Math. Sci.* **24** (1988), 821–847.

[Ha78] U. Haagerup, An example of a nonnuclear C^*-algebra, which has the metric approximation property. *Invent. Math.* **50** (1978/79), 279–293.

[Ho07] C. Houdayer, Construction of type II_1 factors with prescribed countable fundamental group. *J. Reine Angew Math.* **634** (2009), 169–207.

[Ho09] C. Houdayer, Strongly solid group factors which are not interpolated free group factors. *Math. Ann.* **346** (2010), 969–989.

[HS09] C. Houdayer and D. Shlyakhtenko, Strongly solid II_1 factors with an exotic MASA. *Preprint.* `arXiv:0904.1225`

[Io06] A. Ioana, Rigidity results for wreath product II_1 factors. *J. Funct. Anal.* **252** (2007), 763–791.

[Io08] A. Ioana, Cocycle superrigidity for profinite actions of property (T) groups. *Preprint.* `arXiv:0805.2998`

[Io10] A. Ioana, W^*-superrigidity for Bernoulli actions of property (T) groups. *Preprint.* `arXiv:1002.4595`

[IPP05] A. Ioana, J. Peterson and S. Popa, Amalgamated free products of weakly rigid factors and calculation of their symmetry groups. *Acta Math.* **200** (2008), 85–153.

[Jo00] P. Jolissaint, Haagerup approximation property for finite von Neumann algebras. *J. Operator Theory* **48** (2002), 549–571.

[Jo82] V.F.R. Jones, Index for subfactors. *Invent. Math.* **72** (1983), 1–25.

[Ki06] Y. Kida, Measure equivalence rigidity of the mapping class group. *Ann. of Math.*, to appear. `arXiv:math/0607600`

[Ki09] Y. Kida, Rigidity of amalgamated free products in measure equivalence theory. *Preprint.* `arXiv:0902.2888`

[MNP68] V. Mandrekar, M. Nadkarni and D. Patil, Singular invariant measures on the line. *Studia Math.* **35** (1970), 1–13.

[McD69] D. McDuff, Uncountably many II_1 factors. *Ann. of Math.* **90** (1969), 372–377.

[McD70] D. McDuff, Central sequences and the hyperfinite factor. *Proc. London Math. Soc.* **21** (1970), 443–461.

[MvN36] F.J. Murray and J. von Neumann, On rings of operators. *Ann. Math.* **37** (1936), 116–229.

[MvN43] F.J. Murray and J. von Neumann, Rings of operators IV, *Ann. Math.* **44** (1943), 716–808.

[Oz03] N. Ozawa, Solid von Neumann algebras. *Acta Math.* **192** (2004), 111–117.

[OP03] N. Ozawa and S. Popa, Some prime factorization results for type II_1 factors. *Invent. Math.* **156** (2004), 223–234.

[OP07] N. Ozawa and S. Popa, On a class of II_1 factors with at most one Cartan subalgebra. *Ann. Math.*, to appear. `arXiv:0706.3623`

[OP08] N. Ozawa and S. Popa, On a class of II_1 factors with at most one Cartan subalgebra, II. *Amer. J. Math.*, to appear. `arXiv:0807.4270`

[Pe06] J. Peterson, L^2-rigidity in von Neumann algebras. *Invent. Math.* **175** (2009), 417–433.

[Pe09] J. Peterson, Examples of group actions which are virtually W^*-superrigid. *Preprint.* `arXiv:1002.1745`

[PS09] J. Peterson and T. Sinclair, On cocycle superrigidity for Gaussian actions. *Preprint.* arXiv:0910.3958

[Po86] S. Popa, Correspondences. *INCREST Preprint* **56** (1986). Available at www.math.ucla.edu/∼popa/preprints.html

[Po01] S. Popa, On a class of type II_1 factors with Betti numbers invariants. *Ann. of Math.* **163** (2006), 809–899.

[Po03] S. Popa, Strong rigidity of II_1 factors arising from malleable actions of w-rigid groups, I. *Invent. Math.* **165** (2006), 369–408.

[Po04] S. Popa, Strong rigidity of II_1 factors arising from malleable actions of w-rigid groups, II. *Invent. Math.* **165** (2006), 409–452.

[Po05] S. Popa, Cocycle and orbit equivalence superrigidity for malleable actions of w-rigid groups. *Invent. Math.* **170** (2007), 243–295.

[Po06a] S. Popa, On the superrigidity of malleable actions with spectral gap. *J. Amer. Math. Soc.* **21** (2008), 981–1000.

[Po06b] S. Popa, Deformation and rigidity for group actions and von Neumann algebras. In *Proceedings of the International Congress of Mathematicians (Madrid, 2006)*, Vol. I, European Mathematical Society Publishing House, 2007, p. 445–477.

[PV06] S. Popa and S. Vaes, Strong rigidity of generalized Bernoulli actions and computations of their symmetry groups. *Adv. Math.* **217** (2008), 833–872.

[PV08a] S. Popa and S. Vaes, Actions of $\mathbb{F}_\infty$ whose II_1 factors and orbit equivalence relations have prescribed fundamental group. *J. Amer. Math. Soc.* **23** (2010), 383–403.

[PV08b] S. Popa and S. Vaes, Cocycle and orbit superrigidity for lattices in $\mathrm{SL}(n,\mathbb{R})$ acting on homogeneous spaces. In *Geometry, rigidity and group actions,* Eds. B. Farb and D. Fisher. The University of Chicago Press, to appear. arXiv:0810.3630

[PV08c] S. Popa and S. Vaes, On the fundamental group of II_1 factors and equivalence relations arising from group actions. In *Noncommutative geometry, Proceedings of the Conference in honor of A.Connes' 60th birthday*, to appear. arXiv:0810.0706

[PV09] S. Popa and S. Vaes, Group measure space decomposition of II_1 factors and W^*-superrigidity. *Preprint.* arXiv:0906.2765

[Ra91] F. Rădulescu, The fundamental group of the von Neumann algebra of a free group with infinitely many generators is $\mathbb{R}_+ \setminus \{0\}$. *J. Amer. Math. Soc.* **5** (1992), 517–532.

[Sh05] Y. Shalom, Measurable group theory. In *European Congress of Mathematics,* European Mathematical Society Publishing House, 2005, p. 391–423.

[Si55] I.M. Singer, Automorphisms of finite factors. *Amer. J. Math.* **77** (1955), 117–133.

[Ta73] M. Takesaki, Duality for crossed products and the structure of von Neumann algebras of type III. *Acta Math.* **131** (1973), 249–310.

[To05] A. Törnquist, Orbit equivalence and actions of $\mathbb{F}_n$. *J. Symbolic Logic* **71** (2006), 265–282.

[Va06a] S. Vaes, Rigidity results for Bernoulli actions and their von Neumann algebras (after Sorin Popa). Séminaire Bourbaki, exp. no. 961. *Astérisque* **311** (2007), 237–294.

[Va06b] S. Vaes, Factors of type II_1 without non-trivial finite index subfactors. *Trans. Amer. Math. Soc.* **361** (2009), 2587–2606.

[Va07] S. Vaes, Explicit computations of all finite index bimodules for a family of II_1 factors. *Ann. Sci. École Norm. Sup.* **41** (2008), 743–788.

[Vo89] D.V. Voiculescu, Circular and semicircular systems and free product factors. In *Operator algebras, unitary representations, enveloping algebras, and invariant theory (Paris, 1989),* Progr. Math. **92**, Birkhäuser, Boston, 1990, p. 45–60.

[Vo95] D.V. Voiculescu, The analogues of entropy and of Fisher's information measure in free probability theory, III. *Geom. Funct. Anal.* **6** (1996), 172–199.

[vN49] J. von Neumann, On rings of operators. Reduction theory. *Ann. of Math.* **50** (1949), 401–485.

[Zi79] R.J. Zimmer, Strong rigidity for ergodic actions of semisimple Lie groups. *Ann. of Math.* **112** (1980), 511–529.

[Zi80] R.J. Zimmer, On the cohomology of ergodic actions of semisimple Lie groups and discrete subgroups. *Amer. J. Math.* **103** (1981), 937–951.

[Zi84] R.J. Zimmer, Ergodic theory and semisimple groups. *Monographs in Mathematics* **81**, Birkhäuser Verlag, Basel, 1984.

Section 10

Dynamical Systems and Ordinary Differential Equations

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Green Bundles and Related Topics

Marie-Claude Arnaud*

Abstract

For twist maps of the annulus and Tonelli Hamiltonians, two linear bundles, the Green bundles, are defined along the minimizing orbits.

The link between these Green bundles and different notions as: weak and strong hyperbolicity, estimate of the non-zero Lyapunov exponents, tangent cones to minimizing subsets, is explained.

Various results are deduced from these links: the relationship between the hyperbolicity of the Aubry-Mather sets of the twist maps and the C^1-regularity of their support, the almost everywhere C^1-regularity of the essential invariant curves of the twist maps, the link between the Lyapunov exponents and the angles of the Oseledec bundles of minimizing measures, the fact that C^0-integrability implies C^1-integrability on a dense G_δ-subset.

Mathematics Subject Classification (2010). Primary 37E40, 37J50, 37C40; 70H20; Secondary 70H03 70H05 37D05 37D25

Keywords. Twist maps, Tonelli Hamiltonians, minimizing measures, Aubry-Mather sets, Lyapunov exponents, hyperbolic sets, non uniform hyperbolic measures, C^1-regularity, weak KAM theory, Hamilton-Jacobi

1. Introduction

In the study of twist maps or optical Hamiltonians, mathematicians have studied the orbits that can be found via minimization for a long time: an action is associated with such a dynamical system, and an orbit piece corresponds to a critical point of the action. For example, a way to find periodic orbits is to minimize the action among the periodic arcs (for Hamiltonians) or sequences (for twist maps).

*ANR-07-BLAN-0361, ANR DynNonHyp
Université d'Avignon et des Pays de Vaucluse, EA 2151, Analyse non linéaire et Géométrie, F-84018 Avignon, France. E-mail: Marie-Claude.Arnaud@univ-avignon.fr.

More recently, the existence of some globally minimizing orbits has been proved, i.e. the existence of orbits that minimize the action along all the intervals of time. In the case of twist maps, these orbits are contained in some minimizing sets (i.e. sets filled with minimizing orbits) called Aubry-Mather sets, which were independently discovered in the 80's by S. Aubry & P. Le Daeron and J. Mather. In the case of the so-called Tonelli Hamiltonians, their existence was proved by J. Mather in the 90's when he proved the existence of minimizing measures. In the case of a Tonelli Hamiltonian of a cotangent bundle T^*M, some minimizing sets similar to the Aubry-Mather sets, called Aubry sets, also exist.

To have an idea of what these Aubry-Mather sets may be, let us consider the case of a completely integrable twist map of $\mathbb{A} = \mathbb{T}\times\mathbb{R}$: $f : (q, p) \to (q+d\tau(p), p)$. Then the annulus is foliated by invariant circles $\{p = C\}$, which are the Aubry-Mather sets. If we slightly perturb f, a lot of these invariant curves will persist (this is a consequence of the K.A.M. theorems), but some others will become smaller invariant sets, Cantor sets or periodic orbits; these three kinds of sets are Aubry-Mather sets; in a certain way, they are the ghosts of the initial invariant circles.

In the case of a generic twist map of the annulus, a result due to Patrice Le Calvez states that the majority of Aubry-Mather sets are hyperbolic (see [30]). No such result is known for the Tonelli Hamiltonians.

In the case of twist maps, too, we know that some non-hyperbolic Aubry-Mather sets, the K.A.M. curves, may persist after perturbation.

We can then ask ourselves:

Question 1. *is there a means of distinguishing between the hyperbolic and the non hyperbolic Aubry or Aubry-Mather sets? Is there a means of* seeing *the Lyapunov exponents of a minimizing measure when knowing only the measure and not the dynamic?*

For the twist maps, there are three kinds of Aubry-Mather sets:

- the invariant curves, which are never uniformly hyperbolic;
- the periodic orbits, which may be hyperbolic or non hyperbolic; there is, of course, no way to distinguish between a hyperbolic and a non hyperbolic finite orbit if we only know the orbit;
- the Cantor sets, which may be hyperbolic or non hyperbolic; we will give a way to distinguish between hyperbolic and non hyperbolic Cantor sets.

Hence, in the case of twist maps, we obtain a criterion to decide if an Aubry-Mather set is hyperbolic or not, without knowing the dynamic. To be a little more precise, we define a notion of C^1-regularity for the subsets of a manifold, and we prove that hyperbolicity is equivalent to C^1-irregularity. In the case of Tonelli Hamiltonians, we will see that this result is no longer true, but a partial result subsists.

The main tool to prove this kind of result is what is called the pair (G_+, G_-) of Green bundles that are defined along every minimizing orbit. These Lagrangian bundles were introduced by L. Green in 1958 in [26] for geodesic flows to prove some rigidity results. More precisely, the existence of only one Lagrangian invariant bundle transverse to the vertical (for example one Green bundle) is required in order to obtain some rigidity results in Riemannian geometry.

Then these Green bundles were used to characterize the Anosov geodesic flows (see [16], and [29] for related results). In this article, we will be interested in this kind of more dynamical result.

We will introduce the Green bundles, give their main properties and explain what kind of results were recently obtained through their use. Roughly speaking, the Green bundles are the limits of the successive images of the "verticals". Let us mention that in the 30's, G. Birkhoff already used the images of the verticals to obtain some a priori inequalities for the invariant curves of twist maps, i.e. to obtain some Lipschitz regularity results. We will speak of the relations between Green bundles and hyperbolicity, Green bundles and Lyapunov exponents and Green bundles and regularity.

The main definitive results that we give are:

- the characterization of the weak (strong) hyperbolicity of the minimizing measures of twist maps by the C^1-irregularity of their support (section 5); this gives a way to see hyperbolicity;
- for the minimizing measures, the link between the positive Lyapunov exponents and the angle between the Oseledec bundles (section 3); this relation is specific to the case of a twisting dynamical system and doesn't exist for general dynamical systems; this is a way to see the Lyapunov exponents;
- some regularity results for the invariant graphs that are C^0-Lagrangian when we make some dynamical assumptions (see section 4). Roughly speaking, we will see that a slow dynamic implies some regularity.

Let us give the outline of this article:

- in section 2, we recall some well-known facts concerning the twist maps and the Tonelli Hamiltonians, construct the Green bundles, and prove that uniform hyperbolicity is equivalent to the transversality of the Green bundles;
- in section 3, we characterize the number of zero Lyapunov exponents of a minimizing measure by way of the dimension of the intersection of the two Green bundles, and we give some estimates of the non-zero Lyapunov exponents via the angle between the Oseledec bundles;
- in section 4, we explain the relationship between the Green bundles and some cones that are tangent to the minimizing subsets. We deduce some

regularity results such as: every continuous invariant graph of a twist map is C^1 almost everywhere; every C^0-integrable Tonelli Hamiltonian is C^1-integrable on an invariant dense G_δ-subset;

- in section 5 we give a complete explaination of the case of the Aubry-Mather sets of twist map: if they have no isolated point, their (weak or uniform) hyperbolicity is equivalent to their C^1-irregularity;
- in section 6, we give an overview of weak KAM theory;
- in section 7, we explain the link between the weak KAM solutions and the Green bundles. We deduce that the support of any minimizing measure all of whose Lyapunov exponents are zero is almost everywhere C^1-regular.

Let us mention that twist maps and Tonelli Hamiltonians appear in numerous problems issued from physics: motivated by the restricted 3-body problem, H. Poincaré introduced the twist maps at the end of the 19th century. Moreover, all the mechanical systems, that correspond to a Hamiltonian that is the sum of a kinetic energy and a potential energy are Tonelli Hamiltonian (N-body problems, simple pendulum...).

2. Green Bundles and Uniform Hyperbolicity

In this section, we will define the two Green bundles along locally minimizing orbits and prove that their transversality implies some hyperbolicity.

2.1. Well-known facts for twist maps. A *twist map* of the annulus $\mathbb{A} = \mathbb{T} \times \mathbb{R}$ is a C^1-diffeomorphism $f : (q, p) \to (Q, P)$ of $\mathbb{A}$ that is is isotopic to identity, satisfies the twist condition:

(twist): for any lift $F : \mathbb{R}^2 \to \mathbb{R}^2$ of f, if we write: $F(x, y) = (X, Y)$, then the map: $(x, y) \to (x, X(x, y))$ is a C^1 diffeomorphism; and has a global generating function $S : \mathbb{A} \to \mathbb{R}$ i.e. such that: $PdQ - pdq = dS$.

The inverse of a twist map is a twist map. The twist map f is *positive* if the map $(x, y) \to (x, X(x, y))$ is orientation preserving.

In general , the generating function is expressed in the lifted coordinates (x, X): $s(x, X) = S \circ p(x, y)$ where $p : \mathbb{R}^2 \to \mathbb{A}$ is the covering map. We can associate the action functionals $A_{n,m}$ with this generating function: given a sequence $(x_n, \dots x_m)$ of points of $\mathbb{R}$, its action is defined by: $A_{n,m}(x_n, \dots, x_m) = \sum_{k=n}^{m-1} s(x_k, x_{k+1})$. Then $(x_n, \dots, x_m)$ is the projection of an orbit segment of F on the x-axis if, and only if, it is is a critical point of $A_{n,m}$ restricted to the space of sequences $(z_n, \dots, z_m)$ with fixed endpoints: $z_n = x_n$ and $z_m = x_m$. In this case the corresponding orbit segment is $(x_k, y_k)_{n \le k \le m}$ where $y_k = \frac{\partial s}{\partial X}(x_{k-1}, x_k) = -\frac{\partial s}{\partial x}(x_k, x_{k+1})$. A bi-infinite sequence $(x_k)_{k \in \mathbb{Z}}$ is

(globally) minimizing if for any $[n,m] \subset \mathbb{Z}$, $(x_k)_{n\leq k\leq m}$ minimizes the action with fixed endpoints. In this case, it is the projection of a unique orbit of F, and we usually say that the corresponding orbit of f is minimizing. We say that $(x_k)_{k\in\mathbb{Z}}$ is locally minimizing if, for any $[n,m] \subset \mathbb{Z}$, $(x_k)_{n\leq k\leq m}$ locally minimizes the action with fixed endpoints.

In the 80's, J. Mather and S. Aubry & P. Le Daeron proved the existence of minimizing orbits (see [5], [34]). Moreover, they proved that every such minimizing orbit is contained in an invariant compact Lipschitz graph above a part of $\mathbb{T}$ that is the union of some minimizing orbits. These Lipschitz graphs are called Aubry-Mather sets. A very important property of these Aubry-Mather sets is that the projected dynamic (on $\mathbb{T}$) of the dynamic restricted to one of these Aubry-Mather set is the restriction of an orientation preserving bi-Lipschitz homeomorphism of the circle. Hence, we can associate a rotation number with such an Aubry-Mather set. We don't give a precise definition here because we won't need it, but the reader can find more details in [25].

Let us recall that an invariant probability μ is ergodic if the μ-measure of every invariant subset is 0 or 1. An ergodic Borel probability measure with compact support is said to be minimizing if its support contains only minimizing orbits. Then its support is an Aubry-Mather set.

2.2. Well-known facts for Hamiltonians. We may define Tonelli Hamiltonians on the cotangent bundle of any closed manifold, but to avoid some complications in the choice of the coordinates (via a Riemannian connection), we will assume that the manifold is $\mathbb{T}^d$. Then $\mathbb{T}^d$ is endowed with its usual flat Riemannian metric and we denote by $\pi : (q,p) \in \mathbb{T}^d \times \mathbb{R}^d \to q \in \mathbb{T}^d$ the usual projection. A Tonelli Hamiltonian of $\mathbb{T}^d \times \mathbb{R}^d$ is a C^3 map $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ that is super-linear in the fiber and C^2 strictly convex in the fiber (i.e. the Hessian $H_{p,p}$ is positive definite at every point).

This Hamiltonian defines a Hamiltonian flow (φ_t) on $\mathbb{T}^d \times \mathbb{R}^d$, solution to the Hamilton equations:

$$\dot{q} = \frac{\partial H}{\partial p}(q,p); \quad \dot{p} = -\frac{\partial H}{\partial q}(q,p).$$

Let us point out a nice interpretation of the convexity assumption: if $V(q,p) = \ker D\pi(q,p) \subset T_{(q,p)}(\mathbb{T}^d \times \mathbb{R}^d)$ is the linear vertical, for all small enough t, the image $D\varphi_t V(q,p)$ of a vertical by the linearized flow is a Lagrangian subspace transverse to the vertical, a graph of a symmetric matrix, $s_t(\varphi_t(q,p))$, close to $\frac{1}{t}\frac{\partial^2 H}{\partial p^2}(q,p)$; moreover, as long as these images are transverse to the vertical, the family $(s_t(q,p))$ is decreasing for the natural order of the symmetric matrices.

We can associate its Legendre map $\mathcal{L} : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{T}^d \times \mathbb{R}^d$, defined by: $\mathcal{L}(q,p) = (q, \frac{\partial H}{\partial p}(q,p))$ with such a Tonelli Hamiltonian. This Legendre map is a C^2-diffeomorphism. We can define too the Lagrangian $L : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$

associated with H, defined by:

$$L(q,v) = \max_{p \in T_q^* M} (p.v - H(q,p)) = \mathcal{L}^{-1}(q,v).v - H \circ \mathcal{L}^{-1}(q,v).$$

The function L is then as regular as H is and $\gamma : I \to M$ is the projection of an orbit segment of the Hamiltonian flow of H if, and only if, $(\gamma, \dot\gamma)$ is a solution to the Euler-Lagrange equations associated with L:

$$\dot q = v; \quad \frac{d}{dt}\left(\frac{\partial L}{\partial v}(q,v)\right) = \frac{\partial L}{\partial q}(q,v).$$

In this case, the corresponding orbit for the Euler-Lagrange flow (f_t) is given by: $t \to (\gamma(t), \dot\gamma(t))$ and the corresponding orbit for the Hamiltonian flow is: $t \to \mathcal{L}^{-1}(\gamma(t), \dot\gamma(t))$. Hence, the two flows are conjugated by $\mathcal{L}$.

An arc $\gamma : [a,b] \to \mathbb{T}^d$ gives a solution $(\gamma, \dot\gamma)$ to the Euler-Lagrange equations if and only if it is a critical point of the Lagrangian action: $A(\gamma) = \int_a^b L(\gamma, \dot\gamma)$ restricted to the set of C^1 arcs with fixed endpoints. We say that $\gamma : \mathbb{R} \to \mathbb{T}^d$ is a minimizer if for any segment $[a,b]$, $\gamma_{|[a,b]}$ minimizes the Lagrangian action among the C^1-curves having the same endpoints. We say that γ is a local minimizer if for any segment $[a,b]$, $\gamma_{|[a,b]}$ is a local (for the C^0-topology) minimizer of the action defined on the set of C^1-arcs with fixed endpoints. The corresponding orbits will be called (locally) minimizing orbits. J. Mather proved the existence of minimizing orbits in [35], and the existence of minimizing measures with compact support, which are ergodic invariant Borel probability measures whose support contains only minimizing orbits.

A classical result asserts that a curve $\gamma : [a,b] \to \mathbb{T}^d$ is locally minimizing if, and only if, γ is a solution to the Euler-Lagrange equations and $\mathcal{L}^{-1}(\gamma, \dot\gamma)_{|]a,b[}$ has no conjugate points; two points $x, y \in \mathbb{T}^d \times \mathbb{R}^d$ are conjugate if $t \neq 0$ exists so that $\varphi_t(x) = y$ and $D\varphi_t V(x) \cap V(y) \neq 0$.

A way to obtain a lot of locally minimizing measures and orbits is to use the so-called modified Lagrangians: if η is a closed 1-form of $\mathbb{T}^d$, then $L - \eta$ has the same Euler-Lagrange flow as L but not the same minimizing orbits. Hence, the two flows have the same locally minimizing orbits (because they have the same pairs of conjugate points) but not the same minimizing orbits. Using a lot of cohomologically different 1-forms, one finds a lot of locally minimizing orbits and measures.

2.3. Construction of the Green Bundles, first properties.

There is a canonical way to identify $T_{(q,p)}(\mathbb{T}^d \times \mathbb{R}^d)$ with $\mathbb{R}^d \times \mathbb{R}^d$. In these coordinates, we have $V(q,p) = \{0\} \times \mathbb{R}^d$; a Lagrangian subspace is said to be *horizontal* if it is transverse to the vertical and $h(q,p) = \mathbb{R}^d \times \{0\}$ is such a horizontal subspace. Let us recall that the graph of a linear map $S : h(q,p) \to V(q,p)$ is Lagrangian if, and only if, S is symmetric.

In the tangent space along a locally minimizing orbit, we may define two invariant horizontal Lagrangian bundles, called the Green bundles. These bundles were introduced by L. Green in in 1958 in [26] for geodesic flows. P. Foulon

extended the construction to the Finsler metrics in [21] and G. Contreras & R. Iturriaga built them for any Tonelli Hamiltonian in [14]. The construction for the twist maps of the annulus, and more generally for the twist maps of $\mathbb{T}^d \times \mathbb{R}^d$ is due to M. Bialy & R. MacKay (see [9])

The method is the following. We consider a locally minimizing orbit (x_t) where t is a real number or an integer, and at x_0 we construct the family of Lagrangian subspaces that are the images of the verticals by the linearized dynamical system: $G_t(x_0) = D\varphi_t V(x_{-t})$ where (φ_t) designates the Hamiltonian flow or the positive twist map. Because the orbit is minimizing, it has no conjugate points and then all the $G_t(x_0)$ for $t \neq 0$ are transverse to the vertical. There is a natural partial order for the Lagrangian subspaces that are transverse to the vertical. On the annulus, this order is the usual order between the slopes of the lines. In higher dimensions, it corresponds to the usual order for the symmetric operators associated with Lagrangian subspaces that we have just defined. For this relation, the family $(G_t)_{t>0}$ is a decreasing family of Lagrangian subspaces and $(G_{-t})_{t>0}$ is an increasing family. Moreover, we have: $\forall u, t > 0, G_{-u}(x_0) \leq G_t(x_0)$. We have then an increasing sequence of Lagrangian subspaces bounded from above and a decreasing one bounded from below. We can take the limit and define the two Green bundles:

$$G_-(x_0) = \lim_{t\to+\infty} G_{-t}(x_0); \quad G_+(x_0) = \lim_{t\to+\infty} G_t(x_0).$$

These two bundles are Lagrangian, invariant, transverse to the vertical and satisfy:

$$\forall t > 0, G_{-t}(x_0) < G_-(x_0) \leq G_+(x_0) < G_t(x_0).$$

Being the limits of monotone sequences of continuous bundles, these bundles are semicontinuous (see [1], [3]): G_- is lower semicontinuous and G_+ is upper semicontinuous.

Remark. In the case of a Tonelli Hamiltonian, an orbit has no conjugate points if, and only if, there exists an invariant Lagrangian sub-bundle F along this orbit that is transverse to the vertical. In this case, we have along the orbit: $G_- \leq F \leq G_+$.

2.4. A dynamical criterion and some consequences.

The orbit (x_t) being locally minimizing and relatively compact, there is a classical result that gives a way to prove that some vectors are in one of the two Green bundles:

Proposition 1 (dynamical criterion). *Let $v \in T_{x_0}(\mathbb{T}^d \times \mathbb{R}^d)$. Then:*

- *if $v \notin G_-(x_0)$, then $\lim_{t\to+\infty} \|D\pi \circ D\varphi_t.v\| = +\infty$;*
- *if $v \notin G_+(x_0)$, then $\lim_{t\to+\infty} \|D\pi \circ D\varphi_{-t}.v\| = +\infty$.*

Proof. To prove the first point of the proposition, we express the matrix of $D\varphi_t$ in the global coordinates of $\mathbb{R}^d \times \mathbb{R}^d$. Then, we use a linear symplectic change of coordinates along the orbit so that the "horizontal subspace" becomes G_-. Because G_- is between G_{-1} and G_1, which depend continuously on x, this change of coordinates is bounded. Then in these coordinates the matrix of $D\varphi_t(x_0)$ is:

$$M_t(x_0) = \begin{pmatrix} -b_t(x_0)s_{-t}(x_0) & b_t(x_0) \\ 0 & s_t(x_t)b_t(x_0) \end{pmatrix}$$

where $G_t(x)$, which is Lagrangian, is the graph of the symmetric matrix $s_t(x)$. The matrix being symplectic, we have: ${}^t b_t(x_0)s_t(x_t)b_t(x_0) = -(s_{-t}(x_0))^{-1}$. As $(s_t(x_t))_{t\geq 1}$ is bounded by s_{-1} and s_1 and as $(s_{-t}(x_0))_{t>0}$ tends to 0 from below when t tends to $+\infty$, we deduce that $\lim_{t\to\infty} m(b_t) = +\infty$ where $m(b) = \|b^{-1}\|^{-1}$ designates the conorm of b. From this we deduce immediately that if a vector $v = (v_1, v_2)$ is not in G_-, i.e. if $v_2 \neq 0$, then: $\lim_{t\to+\infty} \|D\pi \circ D\varphi_t.v\| = +\infty$. □

Remark.

1) We deduce from the dynamical criterion that in the Hamiltonian case, the Hamiltonian vector-field X_H belongs to the two Green bundles. Because these two Green bundles are Lagrangian, this implies that G_+ and G_- are tangent to the Hamiltonian levels $\{H = c\}$.

2) Moreover, we deduce, too, that if there is an Oseledec splitting (this will be defined precisely in section 3), $T(\mathbb{T}^d \times \mathbb{R}^d) = E^s \oplus E^c \oplus E^u$ above a invariant compact set K without conjugate points, then $E^s \subset G_-$ and $E^u \subset G_+$. Because the flow is symplectic, E^u and E^s are isotropic and orthogonal to E^c for the symplectic form (see [10]). We deduce that $G_- \subset E^s \oplus E^c$, $G_+ \subset E^u \oplus E^c$ and then $G_- \cap G_+ \subset E^c$. Hence, $G_- \cap G_+$, being an isotropic subspace of the symplectic subspace E^c, we obtain: $\dim E^c \geq 2\dim(G_- \cap G_+)$. The dimension of the intersection of the two Green bundles gives a lower bound of the number of zero Lyapunov exponents. We will soon prove that this inequality is, in fact, an equality.

   We have the same results for a hyperbolic or partially hyperbolic dynamic. Let us notice that in the hyperbolic case, G_- (resp. G_+) is nothing else but the stable (resp. unstable) bundle E^s (resp. E^u).

3) Let us consider the case of a K.A.M. torus that is a graph: the dynamic on this torus is C^1 conjugated to a flow of irrational translations on the torus $\mathbb{T}^d$; M. Herman proved in [28] that such a torus is Lagrangian, and it is well-known that any invariant Lagrangian graph is locally minimizing. Then the orbit of every vector tangent to the K.A.M. torus is bounded, and belongs to $G_- \cap G_+$. In this case, the two Green bundles are equal to the tangent space of the invariant torus.

The dynamical criterion is the key argument for proofs of hyperbolicity results. In [16] , [21], [14], the authors prove that if there is no conjugate points in a whole energy level and if the Green bundles are transverse in the tangent space of the energy level, then the flow restricted to this energy level is Anosov. In fact, we may extend these results to the locally minimizing subsets. Let us recall: a subset $K \subset \mathbb{A}$ that is invariant by a twist map f is *hyperbolic* if along K there is an Df-invariant splitting $T_x\mathbb{A} = E^s \oplus E^u$ so that along the stable bundle E^s, Df is uniformly contracting and along the unstable bundle E^u, Df is uniformly expanding. A subset $K \subset \mathbb{T}^d \times \mathbb{R}^d$ is *partially hyperbolic* for the Tonelli flow (φ_t) if there is an invariant splitting $E^s \oplus E^c \oplus E^u$ such that $D\varphi_{t|E^s}$ is uniformly contracting, $D\varphi_{t|E^u}$ is uniformly expanding and $D\varphi_{t|E^c}$ is less contracting than $D\varphi_{t|E^s}$ and less expanding than $D\varphi_{t|E^u}$.

For an Hamiltonian flow, as the flow direction and the energy direction are in E^c, we always have: $\dim E^c \geq 2$.

Theorem 2 (Green bundles and uniform hyperbolicity). *Let K be a compact invariant locally minimizing set. Then:*

- *in the case of a twist map, K is hyperbolic if, and only if, at all points of K, G_+ and G_- are transverse;*
- *in the case of a Hamiltonian flow, if K contains no singularity, K is partially hyperbolic with a center bundle with dimension 2 if, and only if, at all points of K, G_- and G_+ are transverse in the energy level.*

Proof. Let us outline the ideas of the proof in the direct sense in the case of transversality of the Green bundles (for example for twist maps). In the Hamiltonian case, where these two bundles are not transverse in the whole tangent space, we restrict and reduce the dynamic to obtain a symplectic cocycle on $T\mathcal{E}/\mathbb{R}X_H$ where $\mathcal{E}$ designates the energy level: for this symplectic reduced cocycle, there exist two reduced Green bundles that are transverse (see [2]). Hence, we only have to prove the result for transverse Green bundles and symplectic cocycles.

In this case, the dynamical criterion implies that the cocycle is *quasi-hyperbolic*, i.e. that the orbit of any non null vector under the cocycle is unbounded. Quasi-hyperbolic dynamics (or more precisely quasi-Anosov dynamics) were studied by R. Mañé in [31]. Quasi-Anosov dynamics that are not Anosov exist (see [22], [38]), but we proved in [3] that a quasi-hyperbolic symplectic cocycle is hyperbolic. The proof mainly uses the original ideas of Mañé. □

Remark. We have seen that in the hyperbolic case, the Green bundles G_-, G_+ are equal to the stable/unstable bundles E^s, E^u. We have seen, too, that along a KAM curve, the two Green bundles are equal. We deduce from these remarks and from the fact that the two Green bundles are semicontinuous that: if T is a KAM curve and if $\varepsilon > 0$ is a positive number, a neighborhood U of T exists so

that along any hyperbolic invariant locally minimizing set K contained in U, the two Oseledec bundles E^s and E^u are ε-close to each other. In section 3, we will give a refinement of these remarks by giving some inequalities between the Lyapunov exponents and the angle between the stable and unstable Oseledec bundles.

3. Non-uniform Hyperbolicity, Estimations of the Lyapunov Exponents of Minimizing Measures

In this section, we will speak of the link between the angle of the two Green bundles and the Lyapunov exponents of a locally minimizing measure.

If K is an invariant subset of a Tonelli flow or twist map denoted by (φ_t), we will say that there is an Oseledec splitting on K if there exist $\lambda_1 < \cdots < \lambda_m$ and an invariant splitting $T(\mathbb{T}^d \times \mathbb{R}^d) = E_1 \oplus \cdots \oplus E_m$ with constant dimensions above K so that:

$$\forall x \in K, \forall i \in [1,m], \forall v \in E_i(x), \lim_{t\to+\infty} \frac{1}{t} \log \|D\varphi_t v\| = \lim_{t\to-\infty} \frac{1}{t} \log \|D\varphi_t v\| = \lambda_i.$$

The Lyapunov exponents are then the λ_i. The stable bundle is $E^s = \bigoplus_{\lambda_i<0} E_i$, the center bundle is $E^c = \bigoplus_{\lambda_i=0} E_i$ and the unstable bundle is $E^u = \bigoplus_{\lambda_i>0} E_i$. The integer $\dim E_i$ is the multiplicity of λ_i. In the symplectic case, 0 always has an even multiplicity, and the Lyapunov exponents λ and $-\lambda$ have the same multiplicity. In the symplectic case, E^s and E^u are isotropic for the symplectic form and E^c is symplectic and orthogonal to $E^s \oplus E^u$ (see [10]).

Oseledec's theorem ([37]) asserts: if μ is an invariant ergodic Borel probability with compact support of (φ_t), then there exists an invariant subset K with full μ-measure so that there exists an Oseledec splitting on K. The corresponding Lyapunov exponents are called the Lyapunov exponents of μ.

In the case of a discrete dynamical system, we say that the measure μ is *weakly hyperbolic* if all its Lyapunov exponents are non zero. If (φ_t) is a symplectic flow, then the multiplicity of the exponent zero is at least 2 (in the directions of the flow and of the energy), and we say that the measure is *weakly hyperbolic* if $\dim E^c = 2$. In this case, the extended stable and unstable bundles are $\tilde{E}^s = E^s \oplus \mathbb{R}X_H$ and $\tilde{E}^u = E^u \oplus \mathbb{R}X_H$ where X_H designates the vector-field.

3.1. The link between the Green bundles and the number of zero Lyapunov exponents.

Let us now consider a minimizing measure. Because we have assumed that we are only looking at ergodic measures, we can associate its Lyapunov exponents with such a measure. Because the dynamic is symplectic, the number of positive exponents is equal to the number

of negative exponents, and the number of zero exponents is even. We obtain in [3] and [2] the following result:

Theorem 3. *Let μ be a minimizing measure. Then μ has exactly 2ρ zero Lyapunov exponents if, and only if, at μ-almost every point, we have:* $\dim(G_- \cap G_+) = \rho$.

Proof. We assume that μ is a minimizing measure. The map $(q,p) \to \dim(G_-(q,p) \cap G_+(q,p))$ being measurable, invariant and μ being ergodic, it is constant almost everywhere. We denote its value by ρ. Let us assume for a while that $\rho = 0$. We use then some coordinates analogous to the ones used in the proof of proposition 1, but our "horizontal bundle" is now G_+. Let us recall that the matrix of the linearized dynamic in these coordinates is:

$$M_t(x_0) = \begin{pmatrix} -b_t(x_0)s_{-t}(x_0) & b_t(x_0) \\ 0 & s_t(x_t)b_t(x_0) \end{pmatrix}$$

We denote by $s_\pm(x)$ the symmetric matrix whose Green bundle $G_\pm$ is the graph. Then $s_+ = 0$ and the matrix s_- is negative definite almost everywhere.

If the time is continuous, i.e. if the dynamical system is a Hamiltonian flow (φ_t), there exists a time $\tau > 0$ such that μ is ergodic for the map φ_τ. We may assume that $\tau = 1$ and from now we work on with the discrete dynamical system $(\varphi_k)_{k\in\mathbb{Z}}$.

Using an Egorov theorem, we find for every $\varepsilon > 0$ a measurable subset J_ε of $\mathbb{T}^d \times \mathbb{R}^d$ and two constants $\beta > \alpha > 0$ so that $\mu(J_\varepsilon) > 1 - \varepsilon$, $(m(b_n))$ tends uniformly to $+\infty$ on J_ε, and: $\forall x \in J_\varepsilon, \beta\mathbf{1} \geq -s_-(x) \geq \alpha\mathbf{1}$ where $\mathbf{1}$ designates the (symmetric) matrix of identity. Then we choose $N \geq 0$ such that $\forall x \in J_\varepsilon, \forall n \geq N, m(b_n(x)) \geq \frac{2}{\alpha}$.

Using the Birkhoff ergodic theorem, we know that for long intervals of time, the orbit piece of almost every point of J_ε comes back into J_ε in a proportion of time bigger than $1 - 2\varepsilon$. We deduce that there is, in such an orbit piece, a proportion bigger than $\frac{1-2\varepsilon}{N}$ of points that belong to J_ε and that correspond to times whose difference is more than N. If $x_0 \in J_\varepsilon$ is a generic point for μ and if we denote by $m_1 < m_2 < \cdots < m_n < \ldots$ the return times in J_ε so that $m_{n+1} - m_n \geq N$, because the term $-b_n(x_0)s_{-n}(x_0)$ is multiplicative, for a big enough n, we find that $m(-b_{m_n}s_{m_n}(x_0))$ is greater than the product of $n \geq \frac{1-2\varepsilon}{N}.m_n$ terms greater than $\alpha.\frac{2}{\alpha} = 2$. Hence, for almost every point of J_ε and for a big enough n, we obtain: $m(-b_{m_n}s_{m_n}(x_0)) \geq \left(2^{\frac{1-2\varepsilon}{N}}\right)^{m_n}$. This implies that μ has at least d Lyapunov exponents greater than $\frac{1-2\varepsilon}{N}\log 2 > 0$ and finishes the proof in this case.

If $\rho \neq 0$, we have seen that μ has at least 2ρ zero exponents. Then, we consider the restricted-reduced cocycle on $(G_+ + G_-)/G_- \cap G_+$ and we prove that it is a symplectic cocycle whose Green bundles are transverse μ almost everywhere. We apply the previous result to find $2(d-\rho)$ positive Lyapunov exponents. □

3.2. Lower and upper bounds for the positive Lyapunov exponents in the Hamiltonian case. In the case of ergodic measures of a geodesic flow with support filled by locally minimizing orbits, i.e. in the case of measures with no conjugate points, A. Freire and R. Mané proved in [23] a nice formula for the sum of the positive exponents (see [21] and[14] too). A slight improvement of this formula gives:

Theorem 4. *Let μ be a Borel probability measure with no conjugate points that is ergodic for the Hamiltonian flow. If G_+ is the graph of $\mathbb{U}$ and G_- the graph of $\mathbb{S}$, the sum of the positive Lyapunov exponents of μ is equal to:*

$$\Lambda_+(\mu) = \frac{1}{2}\int \mathrm{tr}\left(\frac{\partial^2 H}{\partial p^2}(\mathbb{U}-\mathbb{S})\right) d\mu.$$

Hence, we see that the more distant the Green bundles, the greater is the sum of the positive Lyapunov exponents. This gives an upper bound to the positive Lyapunov exponents. A similar statement was given in the (non published) thesis of G. Kniepper.

We can be more precise by introducing a notion of symplectic angle:

Definition. Let F, G be two Lagrangian subspaces of a symplectic linear space (E,ω) endowed with an adapted scalar product. The *greatest angle* between F and G is defined by:

$$\beta(F,G) = \max_{(v,w)\in(F\setminus\{0\})\times(G\setminus\{0\})} \frac{\omega(v,w)}{\|v\|.\|w\|}.$$

This angle is equal to 0 if, and only if, $F = G$. Otherwise, it is positive. In particular, for a weakly hyperbolic measure, if $\tilde{E}^s = E^s \oplus \mathbb{R}X_H$ and $\tilde{E}^u = E^u \oplus \mathbb{R}X_H$ are the enlarged stable and unstable bundles of the Oseledec splitting, their greatest angle is positive.

To obtain some precise estimates of $\Lambda_+(\mu)$ by using this angle, we need some notations:

Notations. If $C > 0$ is a real number and $K \subset \mathbb{T}^d \times \mathbb{R}^d$ is a compact subset, we denote by $\mathcal{H}_C(K)$ the set of Tonelli Hamiltonians such that:

$$\forall x \in K, \exists t, u \in]0,1], s_t(x) \leq C\mathbf{1} \quad \text{and} \quad s_{-u}(x) \geq -C\mathbf{1};$$

where s_t is the matrix of G_t and $\mathbf{1}$ denotes the matrix of identity.

Hence, we say that the elements of $\mathcal{H}_C(K)$ have a minimal twist of the vertical in K.

An easy consequence of theorem 4 is a formulation using the greatest angle:

Corollary 5. *Let $H \in \mathcal{H}_C(K)$ be a Tonelli Hamiltonian and let μ be a weakly hyperbolic ergodic Borel probability measure whose support in contained in K*

and has no conjugate points. Then:

$$\frac{1}{2}\int m\left(\frac{\partial^2 H}{\partial p^2}\right)\beta(\tilde{E}^s,\tilde{E}^u)d\mu \leq \Lambda_+(\mu) \leq \frac{d(1+C^2)}{2}\int \mathrm{Tr}\left(\frac{\partial^2 H}{\partial p^2}\right).\beta(\tilde{E}^s,\tilde{E}^u)d\mu$$

where $\mathrm{Tr}(b)$ *designates the trace of* b *and* $m(b) = \|b^{-1}\|^{-1}$ *the conorm of* b.

Proof. A consequence of the linearized Hamilton equations is that if the graph $\mathcal{G}$ of a symmetric matrix G is invariant by the linearized flow, then any infinitesimal orbit $(\delta q, G\delta q)$ satisfies the following equation: $\delta\dot{q} = (\frac{\partial^2 H}{\partial p^2}G + \frac{\partial^2 H}{\partial q \partial p})\delta q$.

Hence, we have: $\frac{d}{dt}\det(D\pi \circ D\varphi_{t|\mathcal{G}}) = \mathrm{tr}(\frac{\partial^2 H}{\partial p^2}G + \frac{\partial^2 H}{\partial q\partial p})\det(D\pi \circ D\varphi_{t|\mathcal{G}})$; we deduce: $\frac{1}{T}\log\det(D\pi \circ D\varphi_{T|\mathcal{G}}(q,p))$

$$= \frac{1}{T}\log\det(D\pi(q,p)_{|\mathcal{G}}) + \frac{1}{T}\int_0^T \mathrm{tr}\left(\frac{\partial^2 H}{\partial p^2}G + \frac{\partial^2 H}{\partial q\partial p}\right)(\varphi_t(q,p))dt.$$

Via ergodic Birkhoff's theorem, we deduce for (q,p) generic that:

$$\liminf_{T\to+\infty}\frac{1}{T}\log\det(D\pi \circ D\varphi_{T|\mathcal{G}}(q,p)) = \int \mathrm{tr}\left(\frac{\partial^2 H}{\partial p^2}G + \frac{\partial^2 H}{\partial q\partial p}\right)d\mu.$$

Moreover, we know that $E^s \subset G_- \subset E^{s\perp} = E^c \oplus E^s$ and that $E^u \subset G_+ \subset E^{u\perp} = E^c \oplus E^u$. Hence, the sum of the Lyapunov exponents of the restricted cocycle $(D\varphi_{t|G_+})$ is exactly $\Lambda_+(\mu)$ and the sum of the Lyapunov exponents of the restricted cocycle $(D\varphi_{t|G_-})$ is $\Lambda_-(\mu) = -\Lambda_+(\mu)$. Then we have:

$$\Lambda_+(\mu) = \int \mathrm{tr}\left(\frac{\partial^2 H}{\partial p^2}\mathbb{U} + \frac{\partial^2 H}{\partial q\partial p}\right)d\mu \quad \text{and} \quad -\Lambda_+(\mu) = \int \mathrm{tr}\left(\frac{\partial^2 H}{\partial p^2}\mathbb{S} + \frac{\partial^2 H}{\partial q\partial p}\right)d\mu$$

We obtain the conclusion by subtracting the two equalities. □

We obtain, too, a lower bound for the smallest positive Lyapunov exponent. To explain this, we need a notation:

Notations. If S is a positive semidefinite matrix, then $q_+(S)$ is its smallest positive eigenvalue.

Theorem 6. *Let* μ *be a measure with no conjugate points and with at least one non zero Lyapunov exponent; then its smallest positive Lyapunov exponent* $\lambda(\mu)$ *satisfies:* $\lambda(\mu) \geq \frac{1}{2}\int m(\frac{\partial^2 H}{\partial p^2}).q_+(\mathbb{U}-\mathbb{S})d\mu$.

Hence, the gap between the two Green bundles gives a lower bound of the smallest positive Lyapunov exponent. It is not surprising that when E^s and E^u collapse, the Lyapunov exponents are 0. What is more surprising and specific to the case of Tonelli Hamiltonians is the fact that the bigger the gap between E^s and E^u is , the greater the Lyapunov exponents are: in general, along a hyperbolic orbit, you may have a big angle between the Oseledec bundles and some very small Lyapunov exponents.

Proof. Let μ be an ergodic Borel probability measure with no conjugate points; its support K is compact and H belongs to some $\mathcal{H}_C(K)$. We choose a point (q,p) that is generic for μ and $(\delta q, \mathbb{U}\delta q)$ in the Oseledec bundle corresponding to the smallest positive Lyapunov exponent $\lambda(\mu)$ of μ. Using the linearized Hamilton equations, we obtain:

$$\frac{d}{dt}((\delta q(\mathbb{U}-\mathbb{S})\delta q) = \delta q(\mathbb{U}-\mathbb{S})\frac{\partial^2 H}{\partial p^2}(q_t,p_t)(\mathbb{U}-\mathbb{S})\delta q.$$

Hence:

$$\frac{d}{dt}((\delta q(\mathbb{U}-\mathbb{S})\delta q) \geq m\left(\frac{\partial^2 H}{\partial p^2}\right) q_+(\mathbb{U}-\mathbb{S})\delta q(\mathbb{U}-\mathbb{S})\delta q;$$

and: $\frac{2}{T}\log(\|\delta q(T)\|) + \frac{\log 2C}{T} \geq \frac{1}{T}\log(\delta q(T)(\mathbb{U}-\mathbb{S})(q_T,p_T)\delta q(T)) \geq$

$$\frac{1}{T}\log(\delta q(0)(\mathbb{U}-\mathbb{S})(q,p)\delta q(0)) + \frac{1}{T}\int_0^T m\left(\frac{\partial^2 H}{\partial p^2}(q_t,p_t)\right) q_+((\mathbb{U}-\mathbb{S})(q_t,p_t))dt.$$

Using Birkhoff's ergodic theorem, we obtain:

$$\lambda(\mu) \geq \frac{1}{2}\int m\left(\frac{\partial^2 H}{\partial p^2}\right) q_+(\mathbb{U}-\mathbb{S})d\mu. \qquad \square$$

3.3. The non negative Lyapunov exponent for twist maps.

In this case, we are interested in exactly one Lyapunov exponent. Hence, a formula giving the sum of the positive Lyapunov exponents is enough to bound the unique non negative Lyapunov exponent from below and above. Using the standard coordinates of $\mathbb{A}$, we obtain:

Theorem 7. *Let $f : \mathbb{A} \to \mathbb{A}$ be a positive twist map and let μ be a minimizing measure whose non negative Lyapunov exponent is λ. If s_-, s_+ designate the slopes of the two Green bundles and s_k designates the slope of G_k, we have:*

$$\lambda = \frac{1}{2}\int \log\left(\frac{s_+ - s_{-1}}{s_- - s_{-1}}\right) d\mu = \frac{1}{2}\int \log\left(1 + \frac{s_+ - s_-}{s_- - s_{-1}}\right) d\mu.$$

As in the Hamiltonian case, we see that the greatest/smallest exponent depends only on the deviation of the vertical (s_{-1} in the discrete case, $\frac{\partial^2 H}{\partial p^2}$ and C in the Hamiltonian case) and on the "angle" between the two Green bundles.

Proof. As in the proof of proposition 1, we use coordinates such that G_- is the horizontal bundle; the matrix of Df^n at x is then:

$$M = \begin{pmatrix} b_n(x)(s_-(x) - s_{-n}(x)) & b_n(x) \\ 0 & b_n(x)(s_n(f^n x) - s_-(f^n x)) \end{pmatrix}$$

We know that $G_- \subset E^s \oplus E^c$, hence along G_- we see the Lyapunov exponent $-\lambda$. The entry $b_n(x)(s_-(x) - s_{-n}(x)$, which represents the linearized dynamic along G_-, being multiplicative, we have along any μ-generic orbit: $-\lambda =$

$$\lim_{n\to\infty} \frac{1}{n} \log \left(b_n(x)(s_-(x) - s_{-n}(x)\right)$$
$$= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \log \left(b_1(f^k x)(s_-(f^k x) - s_{-1}(f^k x))\right)$$

In the same way, we have: $\lambda =$

$$\lim_{n\to\infty} \frac{1}{n} \log \left(b_n(x)(s_+(x) - s_{-n}(x)\right)$$
$$= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \log \left(b_1(f^k x)(s_+(f^k x) - s_{-1}(f^k x))\right)$$

By subtracting these two equalities and using Birkhoff's ergodic theorem, we obtain the equality of the theorem. □

4. Invariant Lagrangian Graphs: Problems of C^1-regularity

In this section, we will explain the relation between the cones that are tangent to minimizing sets and the Green bundles. Then we will deduce some regularity results. In particular, we will improve Birkhoff's regularity result that asserts that any essential invariant curve of a twist map is a Lipschitz graph, and we will prove that C^0 integrability implies C^1 integrability on a dense G_δ-subset.

4.1. Two notions of C^1 regularity.

We define what the "tangent vectors" to a set that is not necessarily a manifold are. This definition was given by G. Bouligand in the 30's in [12].

Definition. Let $K \subset \mathbb{T}^d \times \mathbb{R}^d$ be a subset of $\mathbb{T}^d \times \mathbb{R}^d$ and let $x \in K$ be a point of K. We say that $v \in T_x(\mathbb{T}^d \times \mathbb{R}^d)$ belongs to the contingent cone to K at x if there exist a sequence (x_n) of points of K converging to x and a sequence (t_n) of positive real numbers so that: $v = \lim_{n\to\infty} t_n(x_n - x)$. The contingent cone to K at x is denoted by $\mathcal{C}_x(K)$.

We say that $v \in T_x(\mathbb{T}^d \times \mathbb{R}^d)$ belongs to the paratingent cone to K at x if two sequences (x_n) and (y_n) of points of K converging to x and a sequence (t_n) of positive real numbers exist so that: $v = \lim_{n\to\infty} t_n(x_n - y_n)$. The paratingent cone to K at x is denoted by $\mathcal{P}_x(K)$.

As the sets in which we are interested are contained in some Lagrangian graphs, we give the following definitions of C^1 regularity:

Definition. We say that $K \subset \mathbb{T}^d \times \mathbb{R}^d$ is strongly C^1-regular at $x \in K$ if the paratingent cone $\mathcal{P}_x(K)$ is contained in a Lagrangian subspace.

We say that $K \subset \mathbb{T}^d \times \mathbb{R}^d$ is weakly C^1-regular at $x \in K$ if a Lagrangian subspace P of $T_x(\mathbb{T}^d \times \mathbb{R}^d)$ exists so that, for every sequence (x_n) of points of K converging to x, we have: $\limsup_{n\to\infty} \mathcal{C}_{x_n}(K) \subset P$ where the lim sup is taken for the Hausdorff metric. The union of such limits will be called the generalized contingent cone and will be denoted by $\mathcal{C}^*_x(K)$.

Because $\mathcal{C}^*_x(K) \subset \mathcal{P}_x(K)$, strong C^1-regularity implies weak C^1-regularity. Moreover, if K is the graph of a Lipschitz map η above the zero section, then K is weakly C^1-regular if, and only if, K is strongly C^1-regular if, and only if, the map η is C^1.

4.2. Tangent cones and Green bundles. We recalled before that if $x \in \mathbb{T}^d \times \mathbb{R}^d$ is a point, there is a natural order between the Lagrangian subspaces of $T_x(\mathbb{T}^d \times \mathbb{R}^d)$ that are transverse to the vertical. But the tangent (contingent or paratingent) cone to a set is not necessarily contained in a linear Lagrangian subspace. We need, then, a new order to compare such a tangent cone to the Green bundles.

Definition. If $A \subset T_x(\mathbb{T}^d \times \mathbb{R}^d)$ is a subset of $T_x(\mathbb{T}^d \times \mathbb{R}^d)$ an if P, P' are two Lagrangian subspaces of $T_x(\mathbb{T}^d \times \mathbb{R}^d)$ that are transverse to the vertical, we say that A is between P and P' and we write $P \leq A \leq P'$ if, for every $a \in A$, there exists a Lagrangian subspace P_A such that $P \leq P_A \leq P'$.

Let us notice that when A is a Lagrangian subspace, $P \leq A \leq P'$ is not ambiguous: it has the same meaning for the two orders. In the case of a twist map of the annulus, $P \leq A \leq P'$ just means that the slope of every element of A is between the slope of every element of P and the slope of every element of P'. Let us notice that: $P \leq A \leq P \Leftrightarrow A \subset P$. We prove this in [3]:

Theorem 8. *Let $f : \mathbb{A} \to \mathbb{A}$ be a positive twist map and K be an Aubry-Mather set. Then:*

$$\forall x \in K, G_-(x) \leq \mathcal{P}_x(K) \leq G_+(x).$$

Similarly, we have (a slightly different version of this is given in see [1]):

Theorem 9. *Let $\mathcal{G}$ be the graph of a C^0 closed form η that is invariant by the Hamiltonian flow of the Tonelli Hamiltonian H. Then:*

$$\forall x \in \mathcal{G}, G_-(x) \leq \mathcal{P}_x(\mathcal{G}) \leq G_+(x).$$

Hence, for the C^0-Lagrangian graphs, the Green bundles give some bounds for the paratingent cones. We will see in section 7 that for the Aubry sets , we

obtain a weaker result (which concerns only the generalized contingent cone); the reason is that in section 7, we use discontinuous Lagrangian graphs, called pseudo-graphs.

4.3. Regularity of invariant C^0-Lagrangian graphs. A classical result asserts that every C^0-Lagrangian invariant graph is locally minimizing. Then, we use two properties of the Green bundles that we found before: the dynamical criterion and the relation between the Green bundles and the paratingent cone to obtain some regularity results for the C^0-Lagrangian graphs (see [1], [4]). Let us mention that an invariant C^0-Lagrangian graph is always Lipschitz (see [19]), but it may happen that a Lipschitz graph is nowhere C^1.

At first, we obtain some results for small dimensions: in this case, the restricted linearized dynamic cannot tend to ∞ on a set with a non zero Lebesgue measure; hence, the two Green bundles are equal and the paratingent cone is a tangent subspace:

Theorem 10. *Let $f : \mathbb{A} \to \mathbb{A}$ be a twist map and let $\gamma : \mathbb{T} \to \mathbb{R}$ be a continuous map whose graph is invariant by f. Then there exists a dense G_δ subset U of $\mathbb{T}$ whose Lebesgue measure is 1 and so that every t of U is a point of differentiability of γ and a point of continuity of γ'. More precisely, the graph of γ is strongly C^1-regular at every point $(t, \gamma(t))$ with $t \in U$.*

This result improves G. Birkhoff's famous result asserting that such an invariant curve is always Lipschitz (see [11]) and proves that some Lipschitz graphs exist that are invariant by no twist map.

Some examples of twist maps exist that have such an invariant curve that is not C^1. But all the known examples have a rational rotation number. Hence, we ask:

Question 2. *Does an example of an invariant curve with an irrational rotation number that is not C^1 exist?*

Theorem 11. *Let $H : \mathbb{T}^2 \times \mathbb{R}^2 \to \mathbb{R}$ be a Tonelli Hamiltonian all of whose singularities are non-degenerate. Let $\mathcal{G}$ be a C^0- Lagrangian graph that is invariant by the Hamiltonian flow of H. If $\mathcal{G}$ is the graph of λ, then there exists a dense G_δ subset D of $\mathbb{T}^2$ with full Lebesgue measure so that λ is differentiable on D and its derivative is continuous at every point of D. More precisely, $\mathcal{G}$ is strongly C^1-regular at every $(q, \lambda(q))$ with $q \in D$.*

When we can specify the restricted dynamic in such a way that all the orbits of the restricted linearized dynamic are bounded, we have, too, $G_- = G_+$ and some regularity results:

Theorem 12. *Let $f : \mathbb{A} \to \mathbb{A}$ be a twist map and let $\gamma : \mathbb{T} \to \mathbb{R}$ be a continuous map whose graph is invariant by f. Let us assume that the restriction of f to the graph of γ is bi-Lipschitz conjugated to a rotation. Then γ is C^1 and the restriction of f to the graph of γ is C^1 conjugated to a rotation.*

Theorem 13. *Let $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Tonelli Hamiltonian and let $\mathcal{G}$ be a C^0-Lagrangian graph that is invariant by the Hamiltonian flow so that the time one flow restricted to $\mathcal{G}$ is bi-Lipschitz conjugated to a translation of $\mathbb{T}^d$. Then the graph $\mathcal{G}$ is C^1.*

Another interesting application of the Green bundles is a description of what happens in the C^0 completely integrable case:

Definition. Let $f : \mathbb{A} \to \mathbb{A}$ be a twist map. Then f is C^0-*integrable* if $\mathbb{A} = \bigcup_{\gamma \in \Gamma} G(\gamma)$ where:

1. Γ is a subset of $C^0(\mathbb{T}, \mathbb{R})$ and $G(\gamma)$ is the graph of γ;
2. $\forall \gamma_1, \gamma_2 \in \Gamma, \gamma_1 \neq \gamma_2 \Rightarrow G(\gamma_1) \cap G(\gamma_2) = \emptyset$;
3. $\forall \gamma \in \Gamma, f(G(\gamma)) = G(\gamma)$.

Remark. The general reference for this remark is [27].

A theorem of Birkhoff states that under the hypothesis of this definition, every $\gamma \in C^0(\mathbb{T}, \mathbb{R})$ whose graph is invariant by f is Lipschitz and that the set $\mathcal{L}(f)$ of those invariant graphs is closed for the C^0-topology.

If we fix a lift $\tilde{f}$ of f, we can associate with every $\gamma \in \mathcal{L}(f)$ its rotation number $\rho(\gamma)$. Then, if $\gamma_1, \gamma_2 \in \mathcal{L}(f)$, we have: $G(\gamma_1) \cap G(\gamma_2) \neq \emptyset \Rightarrow \rho(\gamma_1) = \rho(\gamma_2)$ and $G(\gamma_1) \cap G(\gamma_2) = \emptyset \Rightarrow \rho(\gamma_1) \neq \rho(\gamma_2)$. We deduce that $\mathcal{L}(f) = \Gamma$ and therefore Γ is closed for the C^0 topology.

Theorem 14. *Let $f : \mathbb{A} \to \mathbb{A}$ be a twist map that is C^0 integrable. Let Γ be the set of $\gamma \in C^0(\mathbb{T}, \mathbb{R})$ whose graph is invariant under f. Then a dense G_δ subset $\mathcal{G}$ of Γ endowed with the C^0-topology exists so that: every $\gamma \in \mathcal{G}$ is C^1. Moreover, in $\mathcal{G}$, the C^0-topology is equal to the C^1-topology.*

We say that a Hamiltonian $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ is C^0-integrable if there exists a partition $\mathcal{P}$ of $\mathbb{T}^d \times \mathbb{R}^d$ into C^0-Lagrangian graphs that are invariant by the flow and so that the map sending an element of $\mathcal{P}$ on its cohomology class sends $\mathcal{P}$ onto $H^1(M, \mathbb{R})$.

Theorem 15. *Let $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ be a C^0-integrable Tonelli Hamiltonian and let Λ_1 be the set of closed 1-forms whose graphs are elements of the partition. Then a dense G_δ subset of Λ_1 all of whose elements are C^1, exists.*

Proof. To prove the two last theorems, we begin by noticing that there is a dense set of completely periodic invariant Lagrangian graphs $\mathcal{G}$, i.e. so that there exists $T > 0$ satisfying: $\varphi_{T|\mathcal{G}} = Id_{\mathcal{G}}$. Using the dynamical criterion, on such a graph we have then $G_- = G_+$. Using the semicontinuity of G_- and G_+, we obtain $G_- = G_+$ on a dense G_δ-subset of invariant graphs, and then these invariant graphs are C^1. □

Question 3. *Are there examples of Tonelli Hamiltonians or twist maps that are C^0-integrable but not C^1-integrable?*

5. The Link Between the Shape of the Aubry-Mather Sets and Their Lyapunov Exponents

In this section, we give a complete characterization of the Aubry-Mather sets that are uniformly hyperbolic and of the minimizing measures with non-zero Lyapunov exponents. The following results are proved in [3].

Theorem 16. *Let f be a twist map and let M be an Aubry-Mather set of f with no isolated point. The two following assertions are equivalent:*

- *for all $x \in M$, M is not strongly C^1-regular at x;*
- *the set M is uniformly hyperbolic (for f).*

An amusing corollary is the following: if M is an Aubry-Mather set with no isolated point for two twist maps f_1 et f_2, then it is uniformly hyperbolic for f_1 if, and only if, it is hyperbolic for f_2.

Proof. For the direct sense, we know that the C^1-irregularity implies the transversality of the two Green bundles and then the uniform hyperbolicity. For the other sense, we prove that when M is uniformly hyperbolic, we have along M: $E^u \cup E^s \subset \mathcal{P}(M)$ and then the irregularity. □

Theorem 17. *Let f be a twist map and let μ be a minimizing measure whose support has no isolated point. The two following assertions are equivalent:*

- *for μ-almost every x, the support of μ, denoted by* $\mathrm{supp}\mu$*, is C^1-regular at x;*
- *the Lyapunov exponents of μ (for f) are zero.*

Proof. The fact that irregularity implies non zero Lyapunov exponents is very similar to what was done in the proof of theorem 16. If now $\mathrm{supp}\mu$ is C^1 regular μ-almost everywhere, the projected dynamic of $f_{|\mathrm{supp}\mu}$ is C^1-conjugated to the initial dynamic $f_{|\mathrm{supp}\mu}$ μ almost everywhere. We can extend this projected dynamic to a bi-Lipschitz homeomorphism h of $\mathbb{T}$. For such a bi-Lipschitz homeomorphism, we may define a kind of modified Lyapunov exponent and prove that it is zero everywhere by using a subtle improvement of Klingman's sub-multiplicative ergodic theorem due to A. Furman (see [24]). □

Hence, knowing the measure μ, we can say if the Lyapunov exponents are zero or not, but the knowledge of its support is a priori not sufficient to deduce

if the Lyapunov exponents are zero or no. To be more precise, it is interesting to answer the following questions:

Question 4. *Do two twist maps f and g and two minimizing measures μ_f for f and μ_g for g exist, so that μ_f and μ_g have the same support but are not equivalent (i.e. not mutually absolutely continuous)?*

Another question concerns the existence of such non-uniformly hyperbolic measures:

Question 5. *Do there exist any minimizing measures with non zero Lyapunov exponents that are not uniformly hyperbolic?*

However, in extreme cases, we obtain a result concerning the link between the support and the Lyapunov exponents:

Corollary 18. *Let f be a twist map and let μ be a minimizing measure whose support has no isolated point. If the support is C^1-regular everywhere, then the Lyapunov exponents of μ are zero.*

It is not hard to see that an Aubry-Mather set is everywhere C^1-regular if, and only if, a C^1 map $\gamma : \mathbb{T} \to \mathbb{R}$ exists, whose graph contains M. In [27], M. Herman gives some examples of Aubry-Mather sets that are invariant by a twist map, contained in a C^1-graph but not contained in an *invariant* continuous curve.

Question 6. *Do any examples of minimizing measures with zero Lyapunov exponents that are not supported in a C^1 curve exist?*

It is possible that the numerical evidence contained in [6] and [7] gives such examples.

6. Weak KAM Theory

In the case of a completely integrable Tonelli Hamiltonian, the manifold $\mathbb{T}^d \times \mathbb{R}^d$ is foliated by invariant Lagrangian tori that are graphs. When we perturb such a Hamiltonian, a lot of these tori persist, due to the strong KAM theorems.

The invariant "pseudographs" that we will study in this section are in a certain sense the ghosts of the invariant Lagrangian graphs. It may happen that a Tonelli Hamiltonian has no invariant Lagrangian graph, but it always has some negatively (resp. positively) invariant discontinuous Lagrangian graphs, called pseudographs, which contain true invariant subsets. This name of "pseudograph" is due to P. Bernard (see [8]) and the proof of the existence of negatively invariant pseudographs is what is called the weak KAM theorem and is due to A. Fathi ([18], [20]). We won't give any proof in this section, but all the results that we give here are proved in [18] or [8].

6.1. The Lax-Oleinik semigroup and its interpretation on pseudographs. Before explaining what a pseudograph is and which kind of transformation of these pseudographs to consider, let us define a semigroup on the set $C^0(\mathbb{T}^d, \mathbb{R})$ of continuous maps from $\mathbb{T}^d$ to $\mathbb{R}$. This semigroup is well-known in PDE.

The *negative semigroup* $(T_t^-)_{t>0}$ *of Lax-Oleinik* is defined on $C^0(\mathbb{T}^d, \mathbb{R})$ by:

$$T_t^- u(q_0) = \inf \left(u(q) + \int_0^t L(\gamma(s), \dot{\gamma}(s)) ds \right);$$

where the infimum is taken on all the C^1 arcs $\gamma : [0, t] \to \mathbb{T}^d$ such that $\gamma(t) = q_0$.

In [18] (see also [8]), A. Fathi proves that for every $\varepsilon > 0$, there exists a constant $K > 0$ such that for every $t \geq \varepsilon$ and every $u \in C^0(\mathbb{T}^d, \mathbb{R})$, the function $T_t^- u$ is K-semi-concave where:

1. A function $v : V \to \mathbb{R}$ defined on a subset V of $\mathbb{R}^d$ is *K-semi-concave* if for every $x \in V$, there exists a linear form p_x defined on $\mathbb{R}^d$ so that:

   $$\forall y \in V, v(y) \leq v(x) + p_x(y - x) + K\|y - x\|^2.$$

   Then we say that p_x is a *K-super-differential* of v at x.

2. Let us fix a finite atlas $\mathcal{A}$ of the manifold $\mathbb{T}^d$; a function $u : \mathbb{T}^d \to \mathbb{R}$ is *K-semi-concave* if for every chart (U, ϕ) belonging to $\mathcal{A}$, $u \circ \phi^{-1}$ is K-semi-concave. Then a *K-super-differential* of u at q is $p_x \circ D\phi(q)$ where p_x is a K-super-differential of $u \circ \phi^{-1}$ at $x = \phi(q)$.

A semi-concave function is always Lipschitz and so differentiable almost everywhere and for such a function, we define its pseudograph: a *pseudograph* is the graph $\mathcal{G}(du)$ of du, where $u : \mathbb{T}^d \to \mathbb{R}$ is a semi-concave function.

As the images of any continuous function by the Lax-Oleinik semigroup are semi-concave, we had better consider the action of the Lax-Oleinik group on the set $\mathcal{SC}(\mathbb{T}^d)$ of the semi-concave functions of $\mathbb{T}^d$.

A very nice interpretation of the action of the semigroup in terms of pseudographs is given in [8]:

Theorem 19. *(P. Bernard) Let $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Tonelli Hamiltonian whose flow is denoted by (φ_t) and let $u : \mathbb{T}^d \to \mathbb{R}$ be a semi-concave function. Then:*

$$\forall t > 0, \overline{\mathcal{G}(dT_t^- u)} \subset \varphi_t(\mathcal{G}(du)).$$

Hence, if we are looking at the pseudographs, the action of T_t^- on $u \in \mathcal{SC}(\mathbb{T}^d)$ consists in cutting the image $\varphi_t(\mathcal{G}(du))$ of the pseudograph of du by the flow, removing some parts of this set to obtain a new pseudograph.

6.2. The weak KAM theorem and Mañé's critical value. The weak KAM theorem, due to A. Fathi, gives us some pseudographs that are invariant by this action:

Theorem 20 (Weak KAM theorem, A. Fathi). *Let $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Tonelli Hamiltonian. There exists a unique constant $c \in \mathbb{R}$ such that the modified semigroup $(\tilde{T}_t^-)$ defined by:*

$$\tilde{T}_t^- u = T_t^- u + ct$$

has at least one fixed point. Such a fixed point is called a weak KAM solution and c is called Mañé's critical value.

We denote by $\mathcal{S}^-(H)$ the set of weak KAM solutions for H. The weak KAM theorem proves the existence of some negatively invariant pseudographs such that: $\forall t > 0, \varphi_{-t}(\overline{\mathcal{G}(du)}) \subset \mathcal{G}(du)$. As said earlier, a compact invariant subset corresponds to such a pseudograph:

$$I(du) = \bigcap_{t>0} \varphi_{-t}(\mathcal{G}(du)) = \bigcap_{t>0} \varphi_{-t}(\overline{\mathcal{G}(du)}).$$

This set $I(du)$ is, in fact, a Lipschitz graph.

There is a relation between the Lax-Oleinik semigroup and the Hamilton-Jacobi equation:

Proposition 21. *(A. Fathi) Let $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Tonelli Hamiltonian and let $u : \mathbb{T}^d \to \mathbb{R}$ be a semi-concave function. Then u is a weak KAM solution associated with H if, and only if, it is a solution to the Hamilton-Jacobi equation $H = c$, i.e. if at every point of differentiability q of u, we have: $H(q, du(q)) = c$.*

Hence, the pseudographs of the weak KAM solutions are all contained in the critical level $\{H = c\}$.

Let us mention that the result given by A. Fathi is more general, because it deals with the viscosity solution of the Hamilton-Jacobi equation; we don't want to define this notion, but the reader can find some material in [15]. A good introduction to the PDE aspects of the weak KAM theory can be found in [17].

6.3. Mather, Aubry and Mañé sets. There is a third characterization of Mañé's critical value c:

$$-c = \inf_{\mu} \int_{\mathbb{T}^d \times \mathbb{R}^d} L d\mu$$

where μ varies among the Borel probability measures on $\mathbb{T}^d \times \mathbb{R}^d$ that are invariant by the Euler-Lagrange flow of L. This lower bound is, in fact, achieved by a measure with compact support. A Borel probability measure μ with compact

support in $\mathbb{T}^d \times \mathbb{R}^d$ is said to be minimizing if it is invariant by the Euler-Lagrange flow and satisfies $-c = \int_{\mathbb{T}^d \times \mathbb{R}^d} L d\mu$. It can be proved that for ergodic measures, this definition is equivalent to the one that we gave before (via the Legendre map). If we denote by $\text{supp}\mu$ the support of the measure μ, the *Mather set* is defined by:

$$\mathcal{M}^*(H) = \bigcup_{\mu} \mathcal{L}^{-1}(\text{supp}\mu).$$

where the intersection is taken on all the minimizing measures.

J. Mather proved that $\mathcal{M}^*(H)$ is an invariant non-empty compact subset of $\mathbb{T}^d \times \mathbb{R}^d$ which is a Lipschitz graph above a compact part of the zero section. A. Fathi proved that the pseudograph of any weak KAM solution contains the Mather set. Moreover, for such a weak KAM solution u, any invariant Borel probability measure whose support is contained in $\mathcal{G}(du)$ is, in fact, the image via the Legendre map of a minimizing measure.

The Aubry set is defined by: $\mathcal{A}^*(H) = \bigcap_{u \in \mathcal{S}^-(H)} I(du)$ and the projected Aubry set is: $\mathcal{A}(H) = \pi(\mathcal{A}^*(H))$. The Aubry set is then an invariant compact Lipschitz graph above a part of the zero section.

The Mañé set is defined by:

$$\mathcal{N}^*(H) = \bigcup_{u \in \mathcal{S}^-(H)} I(du).$$

It is compact and invariant, but in general, it is not a graph.

We have: $\mathcal{M}^*(H) \subset \mathcal{A}^*(H) \subset \mathcal{N}^*(H) \subset \mathcal{E} = H^{-1}(c)$.

There are some other characterizations of the Aubry and Mañé sets (see [32] and [13]). Following Mañé, let us define the Mañé potential. For all $(q_1, q_2) \in M^2$ and all $t > 0$, we define: $a_t(q_1, q_2) = \inf \int_0^t L(\gamma(s), \dot{\gamma}(s)) ds$ where the infimum is taken on all the C^1 curves $\gamma : [0, t] \to \mathbb{T}^d$ such that $\gamma(0) = q_1$ and $\gamma(t) = q_2$. The Mañé potential is defined by: $\Phi(q_1, q_2) = \inf_{t>0}(a_t(q_1, q_2) + ct)$. A curve $\gamma : I \to \mathbb{T}^d$ is said to be semi-static if for all $t_1 < t_2$ in I: $\int_{t_1}^{t_2}(L(\gamma(t), \dot{\gamma}(t)) + c)dt = \Phi(\gamma(t_1), \gamma(t_2))$. Then a curve $\gamma : \mathbb{R} \to \mathbb{T}^d$ is semi-static if, and only if, $\mathcal{L}^{-1}(\gamma, \dot{\gamma})$ is the orbit of a point of the Mañé set.

Following Mather (see [35]), we define the Peierls barrier: $h : \mathbb{T}^d \times \mathbb{T}^d \to \mathbb{R}$ by: $h(q_1, q_2) = \liminf_{t \to +\infty}(a_t(q_1, q_2) + ct)$. A. Fathi proved that, in fact, we have a true (and uniform) limit. Then a point q is in the projected Aubry set if, and only if, $h(q, q) = 0$. Moreover, if $\gamma_n : [0, t_n] \to \mathbb{T}^d$ is a sequence of arcs so that $\gamma_n(0) = \gamma_n(t_n) = q$, $\lim_{n\to\infty} t_n = +\infty$ and $\lim_{n\to\infty} \int_0^{t_n} (L(\gamma_n(t), \dot{\gamma}_n(t)) + c) = 0$, then $\lim_{n\to\infty} \mathcal{L}^{-1}(\gamma_n(t_n), \dot{\gamma}_n(t_n)) = \lim_{n\to\infty} \mathcal{L}^{-1}(\gamma_n(0), \dot{\gamma}_n(0)) = (q, p)$ where (q, p) is the point of the Aubry set so that $\pi(q, p) = q$.

6.4. The symmetrical Hamiltonian and the positive Lax-Oleinik semigroup. If $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ is a Tonelli Hamiltonian, its symmetrical Hamiltonian is defined by: $\check{H}(q,p) = H(q,-p)$. Then its associated Lagrangian is defined by: $\check{L}(q,v) = L(q,-v)$. If (φ_t) (resp. $(\check{\varphi}_t)$) is the Hamiltonian flow associated with H (resp. $\check{H}$), if we denote by $i : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{T}^d \times \mathbb{R}^d$ the involution: $i(q,p) = (q,-p)$, then we have: $\check{\varphi}_t(q,p) = i \circ \varphi_{-t} \circ i(q,p)$. Moreover, $\gamma : \mathbb{R} \to \mathbb{T}^d$ is a solution to the Euler-Lagrange equations for L if, and only if, $\check{\gamma} : \mathbb{R} \to \mathbb{T}^d$ defined by $\check{\gamma}(t) = \gamma(-t)$ is a solution to the Euler-Lagrange equations for $\check{L}$ and $\gamma : [a,b] \to \mathbb{T}^d$ minimizes the Lagrangian action of L between $\gamma(a)$ and $\gamma(b)$ if, and only if, $\check{\gamma} : t \in [-b,-a] \to \gamma(-t) \in \mathbb{T}^d$ minimizes the action of $\check{L}$ between $\gamma(b)$ and $\gamma(a)$. From this remark, we deduce the following expression of the negative Lax-Oleinik semigroup of $\check{L}$:

$$\check{T}_t^-u(q) = \inf_\gamma \left(u(\gamma(0)) + \int_0^t L(\gamma(s), \dot{\gamma}(s))ds \right)$$

where the infimum is taken on all the C^1 curves $\gamma : [0,t] \to \mathbb{T}^d$ such that $\gamma(0) = q$.

We then define the positive Lax-Oleinik semigroup for H by: $T_t^+u(q) = -\check{T}_t^-(-u)$. Hence:

$$T_t^+u(q) = \sup_\gamma \left(u(\gamma(t)) - \int_0^t L(\gamma(s), \dot{\gamma}(s))ds \right).$$

Instead of restricting ourselves to the set of semi-concave functions, we now use the set of semi-convex functions (a function $u : \mathbb{T}^d \to \mathbb{R}$ is K-semi-convex if $-u$ is K-semi-concave). The graph of du where u is semi-convex is called an anti-pseudograph, and the anti-pseudograph $\mathcal{G}(du)$ of any fixed point u of the positive Lax-Oleinik semigroup of H satisfies: $\forall t > 0, \overline{\mathcal{G}(du)} \subset \varphi_{-t}(\mathcal{G}(du))$.

Let us notice that H and $\check{H}$ have the same critical value (use the characterization by the minimizing measures) and that $\mathcal{M}^*(H) = i\left(\mathcal{M}^*(\check{H})\right)$ is contained in the pseudograph of any positive weak KAM solution (for H). Moreover, A. Fathi proved that for any negative weak KAM solution u_-, there exists a unique positive weak KAM solution u_+ such that $u_{-|\mathcal{M}(H)} = u_{+|\mathcal{M}(H)}$. Such a pair (u_-, u_+) of weak KAM solutions with $u_{-|\mathcal{M}(H)} = u_{+|\mathcal{M}(H)}$ is called a pair of conjugate weak KAM solutions. We always have:

$$u_+ \leq u_-; \quad \pi(I(du_-)) = \{q \in M; u_-(q) = u_+(q)\};$$

$$du_{-|\pi(I(du_-))(H)} = du_{+|\pi(I(du_-))}.$$

A consequence is that, for every pair (u_-, u_+) of conjugate weak KAM solutions, the Aubry set (and then the Mather set) of H is in $\mathcal{G}(du_-) \cap \mathcal{G}(du_+)$.

7. Weak KAM Solutions and Green Bundles

We proved in [2] that the Green bundles give some bounds for the "second derivative" of the weak KAM solutions along the Aubry set; what we denote

by $\tilde{G}_\pm$ is a modified Green bundle that is very close to the original Green bundle:

Theorem 22. *Let $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Tonelli Hamiltonian and let (u_-, u_+) be a pair of conjugate weak KAM solutions. Then:*

$$\forall x \in I(du_-), \tilde{G}_-(x) \leq \mathcal{C}_x(\mathcal{G}(du_-)) \leq G_+(x) \quad \text{and} \quad G_-(x) \leq \mathcal{C}_x(\mathcal{G}(du_+)) \leq \tilde{G}_+(x).$$

Proof. The proof is rather technical. We begin by selecting a pseudograph in the images of the physical verticals, where the physical vertical at $x \in \mathbb{T}^d \times \mathbb{R}^d$ is: $\mathcal{V}(x) = \pi^{-1}(\pi(x)) \subset \mathbb{T}^d \times \mathbb{R}^d$. Let $x_0 = (q_0, p_0) \in I(du_-)$. Then for every $t > 0$ we prove that there there exist two C^2- functions $g_t^+, g_t^- : \mathbb{T}^d \to \mathbb{R}$, the first one semi-concave and the second one semi-convex, so that $g_t^-(q_0) = u_-(q_0) = u_+(q_0) = g_t^+(q_0)$, $g_t^- \leq u_+ \leq u_- \leq g_t^+$ and: $\mathcal{G}(dg_t^+) \subset \varphi_t(\mathcal{V}(\varphi_{-t}x))$, $\mathcal{G}(dg_t^-) \subset \varphi_{-t}(\mathcal{V}(\varphi_t x))$. Then we manage to deduce that $\mathcal{C}_x(\mathcal{G}(du_-)) \leq G_t(x)$ and $G_{-t}(x) \leq \mathcal{C}_x(\mathcal{G}(du_+))$ where $G_t(x)$ (resp. $G_{-t}(x)$) is the tangent subspace at x to $\mathcal{G}(dg_t^+)$ (resp. $\mathcal{G}(dg_t^-)$). When t tends to $+\infty$, we find the results of the theorem. □

Corollary 23. *Let $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Tonelli Hamiltonian and let (u_-, u_+) be a pair of conjugate weak KAM solutions. Then:*

$$\forall x \in I(du_-), \tilde{G}_-(x) \leq \mathcal{C}_x(I(du_-)) \leq G_+(x)$$

Using the results of section 3 and the fact that the support of every minimizing measure is contained in a subset $I(du_-)$, we deduce:

Corollary 24. *Let $H : \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Tonelli Hamiltonian and let μ be a minimizing measure all of whose Lyapunov exponents are zero. Then, the support of μ is weakly C^1-regular at μ-almost every point.*

This last result is less complete than the one contained in theorem 17, because we obtain only one implication. In fact, the other implication is not correct in this case: it may happen that a hyperbolic set is very smooth, and then that a hyperbolic measure has a regular support.

References

[1] M.-C. Arnaud, *Fibrés de Green et régularité des graphes C^0-Lagrangiens invariants par un flot de Tonelli*, Ann. Henri Poincaré **9** (2008), no. 5, 881–926.

[2] M.-C. Arnaud, *Green bundles, Lyapunov exponents and regularity along the supports of the minimizing measures* , preprint arXiv:1003.2139

[3] M.-C. Arnaud, *The link between the shape of the Aubry-Mather sets and their Lyapunov exponents*, preprint arXiv:0902.3266

[4] M.-C. Arnaud, *Three results on the regularity of the curves that are invariant by an exact symplectic twist map*, Publ. Math. Inst. Hautes Etudes Sci. No. **109**, 1–17 (2009).

[5] S. Aubry & P. Y. Le Daeron. *The discrete Frenkel-Kontorova model and its extensions.* I. Exact results for the ground-states. Phys. D 8 (1983), no. 3, 381–422.

[6] C. Baesens & R. S. MacKay, *Cantori for multi-harmonic maps*, Physica D **69** (1993) 59–76

[7] C. Baesens & R. S. MacKay, *The one-to-two hole transition for cantori*, Physica D **71** (1994) 372–89

[8] P. Bernard. *The dynamics of pseudographs in convex Hamiltonian systems.* J. Amer. Math. Soc. 21 (2008), no. 3, 615–669.

[9] M. L Bialy & R. S. MacKay, *Symplectic twist maps without conjugate points.* Israel J. Math. **141**, 235–247 (2004).

[10] J. Bochi & M. Viana, *Lyapunov exponents: how frequently are dynamical systems hyperbolic?* Modern dynamical systems and applications, 271–297, Cambridge Univ. Press, Cambridge, 2004.

[11] G. D. Birkhoff, *Sur quelques courbes fermées remarquables*, Bull. Soc. Math. France. 80 (1932), 1–26

[12] G. Bouligand. *Introduction à la géométrie infinitésimale directe* (1932) Librairie Vuibert, Paris.

[13] G. Contreras & R. Iturriaga, *Global minimizers of autonomous Lagrangians*, preprint (2000), 208 pp (enlarged version of 22nd Brazilian Mathematics Colloquium (IMPA), Rio de Janeiro, 1999. 148 pp): www.cimat.mx/ gonzalo/

[14] G. Contreras & R. Iturriaga, *Convex Hamiltonians without conjugate points*, Ergodic Theory Dynam. Systems **19** (1999), no. 4, 901–952.

[15] M. G. Crandall & P. L. Lions, *Viscosity solutions of Hamilton Jacobi equations.* Trans. Amer. Math. Soc. 277 (1983), no. 1, 1–42.

[16] P. Eberlein, *When is a geodesic flow of Anosov type?* I,II. J. Differential Geometry 8 (1973), 437–463; ibid. 8 (1973), 565–577

[17] L. C. Evans, *A survey of partial differential equations methods in weak KAM theory.* Comm. Pure Appl. Math. 57 (2004), no. 4, 445–480.

[18] A. Fathi, *Weak KAM theorems in Lagrangian dynamics*, book in preparation.

[19] A. Fathi, *Regularity of C^1 solutions of the Hamilton-Jacobi equation.* Ann. Fac. Sci. Toulouse Math. (6) **12** no. 4, 479–516 (2003).

[20] A. Fathi, *Théorème KAM faible et théorie de Mather sur les systèmes lagrangiens.* C. R. Acad. Sci. Paris Sér. I Math. 324 (1997), no. 9, 1043–1046.

[21] P. Foulon, *Estimation de l'entropie des systèmes lagrangiens sans points conjugués* Ann. Inst. H. Poincaré Phys. Théor. **57**, no. 2, 117–146 (1992).

[22] J. Franks & C. Robinson, *A quasi-Anosov diffeomorphism that is not Anosov.* Trans. Amer. Math. Soc. 223 (1976), 267–278.

[23] A. Freire & R. Mañé. *On the entropy of the geodesic flow in manifolds without conjugate points.* Invent. Math. 69 (1982), no. 3, 375–392.

[24] A. Furman *On the multiplicative ergodic theorem for uniquely ergodic systems.* Ann. Inst. H. Poincaré Probab. Statist. 33 (1997), no. 6, 797–815.

[25] C. Golé, *Symplectic twist maps. Global variational techniques.* Advanced Series in Nonlinear Dynamics, 18. World Scientific Publishing Co., Inc., River Edge, NJ, 2001. xviii+305

[26] L. W. Green, *A theorem of E. Hopf* Michigan Math. J. **5** 31–34 (1958).

[27] M. Herman, *Sur les courbes invariantes par les difféomorphismes de l'anneau*, Vol. 1, Asterisque **103–104** (1983).

[28] M. Herman, *Inégalités "a priori" pour des tores lagrangiens invariants par des difféomorphismes symplectiques.* Inst. Hautes Etudes Sci. Publ. Math. No. 70 (1989), 47–101 (1990).

[29] W. Klingenberg, *Riemannian manifolds with geodesic flow of Anosov type.* Ann. of Math. (2) 99 (1974), 1–13.

[30] P. Le Calvez, *Les ensembles d'Aubry-Mather d'un difféomorphisme conservatif de l'anneau déviant la verticale sont en général hyperboliques.* C. R. Acad. Sci. Paris Sr. I Math. 306 (1988), no. 1, 51–54.

[31] R. Mañé. *Quasi-Anosov diffeomorphisms and hyperbolic manifolds.* Trans. Amer. Math. Soc. 229 (1977), 351–370.

[32] R. Mañé *Lagrangian flows: the dynamics of globally minimizing orbits.* International Conference on Dynamical Systems (Montevideo, 1995), 120–131, Pitman Res. Notes Math. Ser., 362, Longman, Harlow, 1996.

[33] R. Mañé, *Generic properties and problems of minimizing measures of Lagrangian systems*, Nonlinearity **9** (1996), no. 2, 273–310.

[34] J. N. Mather. *Existence of quasiperiodic orbits for twist homeomorphisms of the annulus.*Topology 21 (1982), no. 4, 457–467.

[35] J. N. Mather. *Action minimizing invariant measures for positive definite Lagrangian systems* Math. Z. 207 (1991), no. 2, 169–207.

[36] J. N. Mather. *Variational construction of connecting orbits.* Ann. Inst. Fourier (Grenoble) 43 (1993), no. 5, 1349–1386.

[37] V. I. Oseledec *A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems.* (Russian) Trudy Moskov. Mat. Obsc. 19 1968 179–210.

[38] C. Robinson, *A quasi-Anosov flow that is not Anosov.* Indiana Univ. Math. J. 25 (1976), no. 8, 763–767.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Arnold's Diffusion: From the *a priori* Unstable to the *a priori* Stable Case

Patrick Bernard*

Abstract

We expose some selected topics concerning the instability of the action variables in *a priori* unstable Hamiltonian systems, and outline a new strategy that may allow to apply these methods to *a priori* stable systems.

Mathematics Subject Classification (2010). 37J40, 37J50, 37C29, 37C50, 37J50.

Keywords. Arnold's diffusion, normally hyperbolic cylinder, partially hyperbolic tori, homoclinic intersections, Weak KAM solutions, variational methods, action minimization.

1. Introduction

A very classical problem in dynamics consists in studying the Hamiltonian system on the symplectic manifolds $T^*\mathbb{T}^n = \mathbb{T}^n \times \mathbb{R}^n$ generated by the Hamiltonian

$$\begin{aligned} H_\epsilon : \mathbb{T} \times T^*\mathbb{T}^n = \mathbb{T} \times \mathbb{T}^n \times \mathbb{R}^n &\longrightarrow \mathbb{R} \\ (t,x,y) &\longmapsto \frac{1}{2}\|y\|^2 + \epsilon G(t,x,y) \end{aligned} \tag{1}$$

where ϵ is a small perturbation parameter. More general unperturbed systems $h(y)$ can be considered instead of $\|y\|^2/2$, but we restrict to that particular case in the present paper in order to simplify some notations. For $\epsilon = 0$, the system is integrable, and the momenta y are integrals of motion. For $\epsilon > 0$, these variables undergo small oscillations. KAM theory implies that these oscillations remain permanently bounded for many initial conditions. For other initial conditions, a large evolution might be possible. By Nekhoroshev theory, it must be extremely

*Membre de l'IUF
CEREMADE, UMR CNRS 7534, Place du Marechal de Lattre de Tassigny, 75775 Paris cedex 16, France. E-mail: patrick.bernard@ceremade.dauphine.fr.

slow. The questions we discuss in the present text is whether this large evolution is actually possible, and to what geometric structures it is associated.

Let us consider a resonant momentum $y_0 = (I_0, 0) \in \mathbb{R}^m \times \mathbb{R}^r = \mathbb{R}^n$, and assume that I_0 is not resonant, which means that $k \cdot I_0$ never belongs to $\mathbb{Z}$ for $k \in \mathbb{Z}^m, k \neq 0$. In order to study the dynamics near the torus $\{y = y_0\}$, it is useful to introduce the notations $x = (\theta, q) \in \mathbb{T}^m \times \mathbb{T}^r$, and $y = (I, p) \in \mathbb{R}^m \times \mathbb{R}^r$, $m + r = n$. In the neighborhood of the torus $\{y = y_0\}$, the dynamics is approximated by the averaged system

$$\frac{1}{2}\|y\|^2 + \epsilon V(q),$$

where

$$V(q) = \int_{\mathbb{T}\times\mathbb{T}^m} G(t, \theta, q, y_0) d\theta dt.$$

Following a classical idea of Poincaré and Arnold, we can try to exploit this observation by considering the system

$$H(t, \theta, q, I, p) = \frac{1}{2}\|p\|^2 + \frac{1}{2}\|I\|^2 - \epsilon V(q) - \mu R(t, \theta, q, I, p) \tag{2}$$

with a second perturbation parameter μ independent from ϵ. We assume that V has a unique non-degenerate minimum, say at $q = 0$. Fixing $\epsilon > 0$, we can study this system for $\mu > 0$ small enough, which is a simpler problem which may give some hints about the dynamics of (1). The reason why instability is more easily proved in (2) than in (1) is the presence of the hyperbolic fixed point at $(0, 0)$ of the (q, p) component of the averaged system. Studying (2) for $\mu > 0$ small enough is thus called the *a priori* unstable problem, or the *a priori* hyperbolic problem. In contrast, the Hamiltonian (1) is called *a priori* stable. The *a priori* unstable case is by now quite well understood for $m = 1$, see [46, 17, 18, 26, 5, 48] for example. The *a priori* unstable case for $m > 1$ and the *a priori* stable case can be considered as widely open, in spite of the important announcements of John Mather in [42]. The starting point in the study of (2) is the famous paper of Arnold, [1]. In this paper, Arnold introduced a particular *a priori* unstable system where some geometric structures associated to diffusion, partially hyperbolic tori (that he called whiskered), their stable and unstable manifolds, and heteroclinic connections, can be almost explicitly described. This geometric structure have been called a transition chain. Most of the subsequent works on the *a priori* unstable problem have consisted in trying to find transition chains in more general cases, but understanding the general *a priori* unstable Hamiltonian have required a change of paradigm: from partially hyperbolic tori to normally hyperbolic cylinders. The variational methods introduced by John Mather in [41] and Ugo Bessi in [9] have also been very influential.

Transforming the understanding gained on the dynamics of (2) to informations on the *a priori* stable case is not an easy task. Since we understand the

system (2) when $m = 1$ the first attempt should be to study (1) in the neighborhood of an $(n-1)$-resonant line, for example the line consisting of momenta of the form $y = (I, 0), I \in \mathbb{R}$. We could hope to prove the existence of drift along such a line by using the *a priori* unstable approximations near each value of y. However, we face the problem that an approximation like (2) holds only in the neighborhood of the torus $\{y = (I, 0)\}$ when the frequency $I \in \mathbb{R}$ is irrational. Near the torus $\{y = (I, 0)\}$ with I rational, one should use an approximation of the form

$$H(t, x, y) = \frac{1}{2}\|y\|^2 - \epsilon W(x) - \mu R(t, x, y)$$

and different methods must be used. This is often called the problem of double resonances when $n = 2$. We will call it the problem of *maximal resonances*.

Our general goal in this paper is to study *a priori* unstable systems with a sufficient generality to be able to gain informations on the *a priori* stable case. We start with a relatively detailed description of the Arnold's example in Section 2, which is also an occasion to settle some notations and introduce some important objects, like the partially hyperbolic tori, their stable and unstable manifolds, and the associated generating functions. Working with these generating functions allows to highlight the connections between the various classical approaches, geometric methods, variational methods, and weak KAM theory. Then, from the end of Section 2 to Section 3, we progressively generalize the setting and indicate how the methods introduced on the example of Arnold can be improved to face the new occurring difficulties. We present the Large Gap Problem, which prevents Arnold's mechanism from being directly applied to general *a priori* unstable systems, and explain how the presence of a normally hyperbolic cylinder can be used to solve this Problem and prove instability in general *a priori* unstable systems. In section 4 we give a new result from [6], on the existence of normally hyperbolic cylinders in the *a priori* stable situation, which should allow to apply the tools exposed in the previous sections to *a priori* stable systems. This suggests a possible strategy to prove the following conjecture:

Conjecture 1. *For a typical perturbation G, there exists two positive numbers ϵ_0 and δ, such that, for each $\epsilon \in]0, \epsilon_0[$, The system (1) has an orbit*

$$(\theta(t), q(t), \dot\theta(t), \dot q(t)) : \mathbb{R} \longrightarrow \mathbb{T} \times \mathbb{T}^{n-1} \times \mathbb{R} \times \mathbb{R}^{n-1}$$

such that $\sup_t \dot\theta - \inf_t \dot\theta > \delta$.

We are currently working on this program in collaboration with Vadim Kaloshin and Ke Zhang. The same conjecture can be stated with a more general unperturbed system $h(y)$, and the same proof should work provided h is convex and smooth. Our strategy of proof does not consist in solving the maximal resonance problem, but rather in observing that the conjectured statement can be reached without solving that difficulty. In that respect, what we expose is much easier than the project of Mather as announced in [42]. The result is

weaker since only limited diffusion is obtained. The maximal resonance problem has to be solved in order to prove the existence of global diffusion on a whole resonant line, or from one resonant line to another. Our strategy, on the other hand, has the advantage of working with all $n \geqslant 2$, while Mather is limited to $n = 2$ at the moment.

2. The Example of Arnold and Some Extensions

Following Arnold [1], we consider the Hamiltonian

$$H(t, \theta, q, I, p) = \frac{1}{2}\|p\|^2 + \frac{1}{2}\|I\|^2 + \epsilon(\cos(2\pi q) - 1)(1 + \mu f(t, \theta, q)) \tag{3}$$

with $(t, \theta, q, I, p) \in \mathbb{T} \times \mathbb{T} \times \mathbb{T} \times \mathbb{R} \times \mathbb{R}$. We will often use the corresponding Lagrangian

$$L(t, \theta, q, \dot{\theta}, \dot{q}) = \frac{1}{2}\|\dot{q}\|^2 + \frac{1}{2}\|\dot{\theta}\|^2 + \epsilon(1 - \cos(2\pi q))(1 + \mu f(t, \theta, q)).$$

We will see $\epsilon > 0$ as a fixed parameter, and discuss mainly the small parameter μ. When $\mu = 0$, the variable I is an integral of motion. Our goal is to study its evolutions for $\mu > 0$. The form of the perturbation is chosen in such a way that the two-dimensional tori

$$\mathcal{T}(a) = \mathbb{T}^2 \times \{0\} \times \{a\} \times \{0\}, \quad a \in \mathbb{R}$$

are invariant in the extended phase space, and carry a linear motion of frequency $(1, a)$. By studying invariant manifolds attached to these invariant tori, Arnold discovered a remarkable diffusion mechanism, now called the Arnold Mechanism, that we are now going to describe. In the case $\mu = 0$, the tori $\mathcal{T}(a)$ appear as the products of the hyperbolic fixed point $\{0, 0\}$ of the pendulum in (q, p) by the invariant torus $\mathbb{T} \times \mathbb{T} \times \{a\}$ of the integrable system in the (t, θ, I) space. They are thus partially hyperbolic, and have stable and unstable manifolds, which coincide and can be given explicitly as

$$\mathcal{W}(a) = \left\{ \left(t, \theta, q, a, \pm\partial_q S_0(q)\right) : (t, \theta, q) \in \mathbb{T}^3 \right\}$$

with

$$S_0(q) = \frac{2\sqrt{\epsilon}}{\pi}(1 - \cos(\pi q)). \tag{4}$$

The coincidence and compactness of these stable manifolds is a very special feature of the unperturbed case $\mu = 0$. For $\mu \neq 0$, the tori $\mathcal{T}(a)$ still have stable and unstable manifolds which can be described as follows: There exists two functions

$$S^{\pm}_{a,\mu}(t, \theta, q) : \mathbb{T} \times \mathbb{T} \times [-3/4, 3/4] \longrightarrow \mathbb{R}, \tag{5}$$

which converge to $\pm S_0$ when $\mu \longrightarrow 0$, and such that the graphs

$$\mathcal{W}^{\pm}_{\mu}(a) = \left\{ \left(t, \theta, q \mod 1, a + \partial_\theta S^{\pm}(t,\theta,q), \partial_q S^{\pm}(t,\theta,q)\right) \right\}$$

are pieces of the stable and unstable manifolds of the torus $\mathcal{T}(a)$. More precisely, the set $\mathcal{W}^+(a)$ is negatively invariant under the extended Hamiltonian flow, and

$$\mathcal{T}(a) = \bigcap_{t \leqslant 0} \varphi^t\big(\mathcal{W}^+(a)\big)$$

while the set $\mathcal{W}^-(a)$ is positively invariant under the extended Hamiltonian flow, and

$$\mathcal{T}(a) = \bigcap_{t \geqslant 0} \varphi^t\big(\mathcal{W}^-(a)\big).$$

The functions $S^{\pm}_a$ solve the Hamilton-Jacobi equation

$$\partial_t S + H(t, \theta, q, a + \partial_\theta S, \partial_q S) = a^2/2,$$

which merely says that the invariant manifolds are contained in the energy level of the torus. The functions $S^{\pm}_a$ have an expression in terms of the action:

$$S^+_a(t,\theta,q) = \int_{-\infty}^{\tau} L(s, \theta^+(s), q^+(s), \dot\theta^+(s), \dot q^+(s)) - a\dot\theta^+(s) + a^2/2 ds \qquad (6)$$
$$S^-_a(t,\theta,q) = \int_{\tau}^{+\infty} L(s, \theta^-(s), q^-(s), \dot\theta^-(s), \dot q^-(s)) - a\dot\theta^-(s) + a^2/2 ds,$$

where τ is any real number such that $\tau \mod 1 = t$, and $(\theta^{\pm}(s), q^{\pm}(s))$ is the solution of the Euler-Lagrange equations such that

$$\theta^{\pm}(\tau) = \theta, q^{\pm}(\tau) = q \mod 1, \dot\theta^{\pm}(\tau) = a + \partial_\theta S^{\pm}(t,\theta,q), \dot q^{\pm}(\tau) = \partial_q S^{\pm}(t,\theta,q).$$

Note that the result does not depend on the choice of τ.

2.1. Homoclinic orbits. If $(T, \Theta, Q) \in \mathbb{T} \times \mathbb{T} \times]1/4, 3/4[$ is a critical point of the function

$$\Delta_a(t,\theta,q) = S^+_a(t,\theta,q) - S^-_a(t,\theta,q-1),$$

then the point

$$\begin{aligned} &\big(T, \Theta, Q \mod 1, a + \partial_\theta S^+_a(T,\Theta,Q), \partial_q S^+_a(T,\Theta,Q)\big) \\ =&\big(T, \Theta, (Q-1) \mod 1, a + \partial_\theta S^-_a(T,\Theta,Q-1), \partial_q S^-_a(T,\Theta,Q-1)\big) \end{aligned}$$

obviously belongs both to $\mathcal{W}^+(a)$ and $\mathcal{W}^-(a)$, hence it is a homoclinic point. It is a transversal homoclinic point if in addition the Hessian of Δ_a has rank two (it can not have rank 3 because the intersection is necessarily one-dimensional). It is not obvious at this point that the function Δ_a necessarily has critical points on the domain $\mathbb{T} \times \mathbb{T} \times]1/4, 3/4[$. When μ is small enough, this follows from:

Lemma 2. *If (T, Q) is a critical point of the function $\bar{\Delta}_a : (t, q) \longmapsto \Delta_a(t, q, 1/2)$, then $(T, Q, 1/2)$ is a critical point of Δ_a, hence the manifolds $\mathcal{W}^-(a)$ and $\mathcal{W}^+(a)$ intersect above $(T, \Theta, 1/2) \in \mathbb{T}^3$. This homoclinic point is transversal if and only if the Hessian of $\bar{\Delta}_a$ at (T, Q) is a non-degenerate 2×2 matrix.*

Note that the function $\bar{\Delta}_a$ is defined on $\mathbb{T}^2$, and therefore it has critical points.

Proof. We have $\partial_t S^+(T, Q, 1/2) = \partial_t S^-(T, Q, -1/2)$, let us denote by e this value. We also have $\partial_\theta S^+(T, Q, 1/2) = \partial_\theta S^-(T, Q, -1/2)$, we denote by I this value. It is enough to prove that $\partial_q S^+(T, Q, 1/2) = \partial_q S^-(T, Q, -1/2)$. In order to do so, it is enough to observe that $\partial_q S^+$ is the only non-negative solution of the equation

$$e + H(T, \Theta, 1/2, a + I, .) = a^2/2,$$

and that precisely the same characterization is true for $\partial_q S^-(T, Q, -1/2)$. Note that the equation above has two solutions, and that we can discriminate between them because we work in a perturbative setting which gives us rough informations on the signs. In more general situation, this is a source of difficulty. □

2.2. Heteroclinic orbits. We have proved the existence of homoclinic orbits. But what is interesting for Arnold diffusion are heteroclinic orbits between different tori. We can deduce the existence of a heteroclinic orbit between $\mathcal{T}(a)$ and $\mathcal{T}(a')$ provided we can find a critical point of the function

$$\mathbb{T} \times \mathbb{R} \times]1/4, 3/4[\ni (t, \theta, q) \longmapsto S_a^+(t, \theta, q) - S_{a'}^-(t, \theta, q - 1) + (a - a')\theta,$$

where we have lifted the functions S without changing their names. As before, we can limit ourselves to finding critical points of the function

$$\Sigma_{a,a'} : \mathbb{T} \times \mathbb{R} \ni (t, \theta) \longmapsto S_a^+(t, \theta, 1/2) - S_{a'}^-(t, \theta, -1/2) + (a - a')\theta, \qquad (7)$$

but the term $(a - a')\theta$ prevents us from finding them using a global variational method when $a' \neq a$. This reflects the fact that we are studying a non exact Lagrangian intersection problem. For $\mu = 0$, heteroclinics do not exist. However, recalling that $\bar{\Delta}_a(t, \theta) = \Delta_a(t, \theta, 1/2)$, we have:

Lemma 3. *If the function $\bar{\Delta}_a(t, q)$ has a non-degenerate critical point, then the functions $\Sigma_{a,a'}$ and $\Sigma_{a',a}$ both have a non-degenerate critical point provided a' is sufficiently close to a.*

Proof. The theory of partial hyperbolicity implies that the stable and unstable manifolds $\mathcal{W}_\mu^\pm(a)$ depend regularly on the parameter a. As a consequence, their generating functions $S_a^\pm$ also regularly depend on a, and the functions $\Sigma_{a,a'}$ depend regularly on a and a'. The result follows since $\Sigma_{a,a} = \bar{\Delta}_a$. □

We say that $a_0, a_1, \dots, a_k$ is an elementary transition chain if the functions Σ_{a_{i-1},a_i} have non-degenerate critical points. We will sometimes use the same terminology for the different requirement that these functions have isolated local minima. From Lemma 3, we deduce:

Proposition 1. *Let μ be given and sufficiently small. Let $[a^-, a^+]$ be an interval such that each of the functions $\bar{\Delta}_{a,\mu}, a \in [a^-, a^+]$ have a non-degenerate critical point, which means that each of the tori $\mathcal{T}_\mu(a), a \in [a^-, a^+]$ has a transversal homoclinic orbit. Then there exists an elementary transition chain $a^- = a_0, a_1, \dots, a_k = a^+$.*

Proof. Let us consider the set $A \subset [a^-, a^+]$ of points that can be reached from a^- by a transition chain. The set A is open : If $a' \in A$, then there exists a transition chain $a^- = a_0, a_1, \dots a_k = a'$ and, by Lemma 3, the sequence $a^- = a_0, a_1, \dots a_k, a_{k+1}$ is a transition chain when a_{k+1} is sufficiently close to a. The set A is closed : Let a be in the closure of A. By Lemma 3, the pair a, a' is a transition chain when a' is close to a. Since a is in the closure of A, the point a' can be taken in A. Then, there exists a transition chain $a^- = a_0, \dots, a_k = a'$, and then the longer sequence $a^- = a_0, \dots, a_k, a_{k+1} = a$ is a transition chain between a_0 and a, hence $a \in A$. Being open, closed and not empty (it contains a_0), the set A is equal to $[a^-, a^+]$. □

The existence of transition chains implies the existence of diffusion orbits. This is proved by Arnold invoking an "obstruction property". This obstruction property is a characteristic of the local dynamics near the partially hyperbolic tori. It has been proved by Jean-Pierre Marco in [38], see also [21, 31]. The most appealing way to understand the geometric shadowing of transition chains is to use the following statement of Jacky Cresson [22], which can be seen as a strong obstruction property:

Lemma 4. *If there exists a transversal heteroclinic between $\mathcal{T}(a)$ and $\mathcal{T}(a')$ and a transversal heteroclinic between $\mathcal{T}(a')$ and $\mathcal{T}(a'')$, then there exists a transversal heteroclinic between $\mathcal{T}(a)$ and $\mathcal{T}(a'')$.*

This Lemma implies:

Corollary 5. *If $a_0, a_1, \dots, a_k$ is an elementary transition chain, then there exists a transversal heteroclinic orbit between $\mathcal{T}(a_0)$ and $\mathcal{T}(a_k)$.*

Putting everything together, we obtain:

Theorem 1. *Let μ be given and sufficiently small. Let $[a^-, a^+]$ be an interval such that each of the functions $\bar{\Delta}_{a,\mu}, a \in [a^-, a^+]$ have a non-degenerate critical point. Then there exists a heteroclinic orbit between $\mathcal{T}(a^-)$ and $\mathcal{T}(a^+)$.*

2.3. Poincaré-Melnikov approximation.

We have constructed diffusion orbits under the assumption that transversal homoclinics exist. We

have proved that homoclinic orbits necessarily exist, and one may argue that transversality should hold for typical systems, we will come back on this later. However, it is useful to be able to check whether transversality holds in a given system. A classical approach consists in proving the existence of non-degenerate critical points of the functions $\bar{\Delta}_{a,\mu}$ defined in Lemma 3 by expanding them in power series of μ. As a starting point the generating functions $S_a^{\pm}$ can be expanded as follows:

$$\begin{aligned} S_a^+(t,\theta,q) &= S_0(q) + \mu M_a^+(t,\theta,q) + O(\mu^2), \\ S_a^-(t,\theta,q) &= -S_0(q) - \mu M_a^-(t,\theta,q) + O(\mu^2), \end{aligned} \tag{8}$$

where $S_0(q) = \frac{2\sqrt{\epsilon}}{\pi}(1-\cos(\pi q))$ is the generating function of the unperturbed manifolds, and $M^{\pm}$ are the so-called Poincaré-Melnikov integrals,

$$M_a^+(t,\theta,q) = \epsilon \int_{-\infty}^{t} F\left(s, \theta + a(s-t), \frac{2}{\pi}\arctan\left(e^{2\pi\sqrt{\epsilon}(s-t)}\tan(\pi q/2)\right)\right) ds$$

$$M_a^-(t,\theta,q) = \epsilon \int_{t}^{+\infty} F\left(s, \theta + a(s-t), \frac{2}{\pi}\arctan\left(e^{2\pi\sqrt{\epsilon}(t-s)}\tan(\pi q/2)\right)\right) ds$$

where $F(t,\theta,q) = (1-\cos(2\pi q))f(t,\theta,q)$. To better understand these formula, it is worth recalling that

$$s \longmapsto \frac{2}{\pi}\arctan\left(e^{2\pi\sqrt{\epsilon}(s-t)}\tan(\pi q/2)\right)$$

is the homoclinic orbit of the system $\|p\|^2/2 + \epsilon(\cos(2\pi q) - 1)$ which takes the value q at time t. The formula above are similar to (6), but the integration is performed on unperturbed trajectories, which are explicitly known. For $q \in]1/4, 3/4[$, we obtain:

$$\Delta_a(t,\theta,q) = S_a^+(t,\theta,q) - S_a^-(t,\theta,q-1) = \mu M_a(t,\theta,q) + O(\mu^2),$$

where M_a is the Poincaré-Melnikov integral

$$\begin{aligned} M_a(t,\theta,q) &= M_a^+(t,\theta,q) + M_a^-(t,\theta,q-1) \\ &= \epsilon \int_{\mathbb{R}} F\left(s, \theta + a(s-t), \frac{2}{\pi}\arctan\left(e^{2\pi\sqrt{\epsilon}(s-t)}\tan(\pi q/2)\right)\right) ds. \end{aligned}$$

In the specific case studied by Arnold, where $f(t,\theta,q) = \cos(2\pi\theta) + \cos(2\pi t)$, the Melnikov integral can be computed explicitly through residues, we obtain:

$$M_a(t,q,1/2) = \frac{a}{\mathrm{sh}(\pi a/2\sqrt{\epsilon})}\cos(2\pi\theta) + \frac{1}{\mathrm{sh}(\pi/2\sqrt{\epsilon})}\cos(2\pi t),$$

it has a non-degenerate minimum at $(t,q) = (0,0)$. We can conclude, following Arnold:

Theorem 2 (Arnold, [1]). *Let us consider the Hamiltonian (3) with $f(t,\theta,q) = \cos(2\pi\theta) + \cos(2\pi t)$ and $\mu > 0$ small enough. Given two real numbers $a^- < a^+$, there exists an orbit $(\theta(t), q(t), I(t), p(t))$ and a time $T > 0$ such that $I(0) \leqslant a^-$ and $I(T) \geqslant a^+$.*

2.4. Bessi's variational mechanism. Ugo Bessi introduced in [9] a very interesting approach to study the system (3), see also [10, 11]. In order to describe this approach, let us define the function

$$A_a : \mathbb{R}\times\mathbb{T}\times]1/4,3/4[\times\mathbb{R}\times\mathbb{T}\times]1/4,3/4[\longrightarrow\mathbb{R}$$
$$((t_1,\theta_1,q_1),(t_2,\theta_2,q_2))\longmapsto \min\int_{t_1}^{t_2} L(s,\theta(s),q(s),\dot\theta(s),\dot q(s))-a\dot\theta(s)+a^2/2\,ds,$$

where the minimum is taken on the set of C^1 curves $(\theta(s),q(s)) : [t_1,t_2]\longrightarrow \mathbb{T}\times\mathbb{R}$ such that

$$(\theta(t_1),q(t_1))=(\theta_1,q_1-1)\quad\text{and}\quad(\theta(t_2),q(t_2))=(\theta_2,q_2).$$

When the time interval t_2-t_1 is very large, the minimizing trajectory in the definition of A_a roughly looks like the concatenation of an orbit positively asymptotic to $\mathcal{T}(a)$ followed by an orbit negatively asymptotic to $\mathcal{T}(a)$. Using this observation, and recalling the formula (6), it is possible to prove that

$$A_a((t_1,\theta_1,q_1),(t_2+k,\theta_2,q_2))\longrightarrow$$
$$S_a^+(t_2\mod 1,\theta_2,q_2)-S_a^-(t_1\mod 1,\theta_1,q_1-1)$$

when $k\longrightarrow\infty$. Fixing the real numbers $a_0,a_1,\dots,a_k$ and the integers $\tau_1,\dots,\tau_k$, we consider the discrete action functional

$$\begin{aligned}
&S_{a_0}^+(t_1\mod 1,\theta_1\mod 1,1/2)+(a_0-a_1)\theta_1\\
&+A_{a_1}((t_1,\theta_1\mod 1,1/2),(t_2+\tau_2,\theta_2\mod 1,1/2))+(a_1-a_2)\theta_2\\
&+A_{a_2}((t_2,\theta_2\mod 1,1/2),(t_3+\tau_3,\theta_3\mod 1,1/2))+(a_2-a_3)\theta_3\\
&+\cdots\\
&+A_{a_{k-1}}(t_{k-1}+\tau_{k-1},\theta_{k-1}\mod 1,1/2),(t_k,\theta_k\mod 1,1/2))+(a_{k-1}-a_k)\theta_k\\
&-S_{a_k}^-(t_k\mod 1,\theta_k\mod 1,1/2)
\end{aligned}$$

defined on $(]-1,1[\times]-1,1[)^k$. It is not hard to check that local minima of this discrete action functional give heteroclinics between the Torus $\mathcal{T}(a_0)$ and the torus $\mathcal{T}(a_k)$. In order to prove that local minima exist, observe that this functional is approximated by

$$\Sigma_{a_0,a_1}(t_1\mod 1,\theta_1)+\cdots+\Sigma_{a_{k-1},a_k}(t_k\mod 1,\theta_k)$$

when the integers τ_i are large enough, with the functions Σ as defined in (7). This limit functional has the remarkable structure that the variables (t_i,θ_i) are separated. This break-down of the action functional into a sum of independent functions is sometimes called an anti-integrable limit, it is related to the obstruction property of the invariant tori, to the λ-Lemma, and to the Shilnokov's Lemma, see [15]. The limit functional has an isolated local minimum provided each of the functions Σ_{a_{i-1},a_i} has one, which is equivalent to say

that $a_0, a_1, \ldots, a_k$ is an elementary transition chain. In this case, the integers τ_i can be chosen large enough so that the action functional above has a local minimum, which gives a heteroclinic orbit between $\mathcal{T}(a_0)$ and $\mathcal{T}(a_k)$. Technically, this method has several advantages. In our presentation we introduced the generating functions $S^{\pm}_{a_i}$ of the invariant manifolds of the involved tori in order to stress the relations between the geometric and the variational method, and also because this works in a more general setting, see [15] for example. In our context and when $\mu > 0$ is small enough, it is easier to directly approximate the functions A_a in terms of the Melnikov integrals, and to use the following approximation for the action functional with large τ_i and small μ without the intermediate step through $S^{\pm}$:

$$\mu M_{a_0}(t_0 \mod 1, \theta_0 \mod 1, 1/2) + (a_1 - a_0)\theta_0 + \cdots + \mu M_{a_k}(t_k \mod 1, \theta_k \mod 1, 1/2) + (a_k - a_{k-1})\theta_k.$$

The corresponding calculations, performed in [9], are much more elementary than those required to derive the expansions (8).

2.5. Remarks on estimates. We have up to that point carefully avoided to discuss the subtle and important aspect of explicit estimates. In order to complete rigorously the proof of Theorem 2, we should prove the existence of a threshold $\mu_0(\epsilon)$ such that the Melnikov approximation holds, simultaneously for all a, when $0 < \mu < \mu_0(\epsilon)$. This can actually been done, with

$$\mu_0(\epsilon) = e^{-\frac{C}{\sqrt{\epsilon}}},$$

but it is not simple, since it requires to study carefully the expansions of the functions $S^{\pm}_a$ and how the coefficients depend on a and ϵ. This is related to the so-called splitting problem, see [37]. As we mentioned above the approach of Bessi allows to prove that Theorem 2 holds for $0 < \mu < \mu_0(\epsilon)$ without estimating the splitting.

It is also important to give time estimates, that is to estimate the time needed for the variable I to perform a large evolution. Once again, this is closely related to the splitting estimates, although these can be avoided by using the method of Bessi. One should distinguish two different problems. Either we fix ϵ, and try to estimate the time as a function of μ, or we take μ as a function of ϵ, say $\mu = \mu_0(\epsilon)/2$, and try to estimate the time as a function of ϵ.

The second problem is especially important, because it is relevant for the study of the *a priori* stable problem. Once again, Ugo Bessi obtained the first estimate,

$$T = e^{\frac{C}{\sqrt{\epsilon}}}.$$

Estimating the time on examples allows to test the optimality of Nekhoroshev exponents, see [39, 36, 49] for works in that direction, see also [16] concerning the question of time estimates.

It is worth mentioning also that in the first problem, estimating the time as a function of μ, the estimate is polynomial, and not exponentially small. This was first understood by Pierre Lochak, and proved by Bessi's method in [2], where the estimate $T = C/\mu^2$ is given, see also [23]. The optimal estimate is $T = C|\ln\mu|/\mu$, as was conjectured by Lochak in [35] and proved by Berti, Biasco and Bolle in [8], see also [7].

Returning to the question of the threshold of validity, let us discuss what happens when μ is increased above $\mu_0(\epsilon)$. The content of Section 2.3 on finding transversal homoclinics via the Poincaré-Melnikov approximation breaks down, but the geometric constructions of the earlier sections is still valid. Theorem 1 holds as long as the invariant tori $\mathcal{T}(a)$ remain partially hyperbolic, and that their stable and unstable manifold can be represented by generating functions like (5). Actually, the methods we are now going to expose allow even to relax this last assumption. Being able to treat larger values of μ is especially important in view of the possible applicability to the *a priori* stable problem.

2.6. Higher dimensions. Let us now discuss the following immediate generalization in higher dimensions of Arnold's example:

$$H(t,\theta,q,I,p) = \frac{1}{2}\|p\|^2 + \frac{1}{2}\|I\|^2 - \epsilon V(q)(1+\mu f(t,\theta,q))$$

with $(t,\theta,q,I,p) \in \mathbb{T}\times\mathbb{T}^m\times\mathbb{T}^r\times\mathbb{R}^m\times\mathbb{R}^r$, where $V(q)$ is a non-negative function having a unique non-degenerate minimum at $q=0$, with $V(0)=0$. The main difference with the example of Arnold appears for $r>1$. In this case, the system is not integrable even for $\mu=0$. There still exists a family of partially hyperbolic tori of dimension m,

$$\mathcal{T}(a) := \{(t,\theta,0,a,0), (t,\theta)\in\mathbb{T}\times\mathbb{T}^m\}$$

parametrized by $a\in\mathbb{R}^m$, but the system $\|p\|^2/2 - \epsilon V(q)$ is not necessarily integrable any more. As a consequence we do not know explicitly the stable and unstable manifolds of the hyperbolic fixed point $(0,0)$, and so we do not have a perturbative setting to describe the stable and unstable manifolds of the hyperbolic tori $\mathcal{T}(a)$. This is also what happens for $r=1$ if μ is not small enough. There is no obvious generalization of the generating functions $S_a^{\pm}$ in that setting, because the stable and unstable manifolds are not necessarily graphs over a prescribed domain. The proof of the existence of homoclinic orbits as given in 2.1 thus breaks down. The existence of homoclinic orbits in that setting can still be proved by global variational methods, as is now quite well understood, see [14, 28, 20, 3, 27] for example.

The proof is quite easy in our context, let us give a rapid sketch. We first define a function A_a similar to the one appearing in Section 2.4, but slightly

different:

$$A_a : \mathbb{R} \times \mathbb{T}^m \times \mathbb{T}^r \times \mathbb{R} \times \mathbb{T}^m \times \mathbb{T}^r \longrightarrow \mathbb{R}$$

$$((t_1, \theta_1, q_1), (t_2, \theta_2, q_2)) \longmapsto \min \int_{t_1}^{t_2} L(s, \theta(s), q(s), \dot{\theta}(s), \dot{q}(s)) - a\dot{\theta}(s) + a^2/2\, ds,$$

where the minimum is taken on the set of curves $(\theta(s), q(s)) : [t_1, t_2] \longrightarrow \mathbb{T}^m \times \mathbb{T}^r$ such that $(\theta(t_i), q(t_i)) = (\theta_i, q_i)$ for $i = 1$ or 2. Let us set

$$\xi(a) := \liminf_{\mathbb{N} \ni k \longrightarrow \infty} A_a((0, 0, q_0), (k, 0, q_1)),$$

and consider a sequence of minimizing extremals

$$(\theta_i(t), q_i(t)) : [0, k_i] \longrightarrow \mathbb{T}^m \times \mathbb{T}^r$$

such that $(\theta_i(0), q_i(0)) = (0, q_0)$, $(\theta_i(k_i), q_i(k_i)) = (0, q_1)$, $k_i \longrightarrow \infty$, and

$$\int_0^{k_i} L(s, \theta_i(s), q_i(s), \dot{\theta}_i(s), \dot{q}_i(s)) - a\dot{\theta}_i(s) + a^2/2\, ds \longrightarrow \xi(a).$$

Let M be a submanifold of $\mathbb{T} \times \mathbb{T}^m \times \mathbb{T}^r$ which separates $\mathbb{T} \times \mathbb{T}^m \times \{q_0\}$ from $\mathbb{T} \times \mathbb{T}^m \times \{q_1\}$, and let $T_i \in [0, k_i]$ be a time such that $(T_i \mod 1, \theta_i(T_i), q_i(T_i)) \in M$, and let τ_i be the integer part of T_i. It is not hard to check that the curves $(\theta_i(t - \tau_i), q_i(t - \tau_i))$ converge (up to a subsequence) uniformly on compact sets to a limit $(\theta_\infty(t), q_\infty(t)) : \mathbb{R} \longrightarrow \mathbb{T}^m \times \mathbb{T}^r$. This limit curve satisfies

$$\int_{-\infty}^{\infty} L(s, \theta_\infty(s), q_\infty(s), \dot{\theta}_\infty(s), \dot{q}_\infty(s)) - a\dot{\theta}_\infty(s) + a^2/2\, ds = \xi(a), \qquad (9)$$

and the corresponding orbit is a heteroclinic from $\mathcal{T}_0(a)$ to $\mathcal{T}_1(a)$. We call minimizing heteroclinics (for the lifted system) those which have minimal action, or in other words those which satisfy (9). In the original system (before taking the covering), we call minimizing homoclinic orbit a homoclinic which lifts to a minimizing heteroclinic.

Let us now try to establish some connections between the present discussion and the proof of the existence of homoclinic orbits given in 2.1. We define two functions on $\mathbb{T} \times \mathbb{T}^m \times \mathbb{T}^r$:

$$S_a^-(t, \theta, q) = -\liminf_{\mathbb{N} \ni k \longrightarrow \infty} A_a\big((t, \theta, q), (k, 0, q_1)\big) \qquad (10)$$

$$S_a^+(t, \theta, q) = \liminf_{\mathbb{N} \ni k \longrightarrow \infty} A_a\big((0, 0, q_0), (t + k, \theta, q)\big). \qquad (11)$$

Note that

$$S_a^+(0, 0, q_1) = -S_a^-(0, 0, q_0) = \xi(a).$$

The functions $S_a^\pm$, whose definition is basic both in Mather's ([41]) and in Fathi's ([29]) theory, share many features with those introduced in (5), that's why we use the same name. Let us state some of their properties:

The function S_a^- is non-positive and it vanishes on $\mathbb{T}\times\mathbb{T}^m\times\{q_1\}$ (and only there). Moreover, it is smooth around this manifold, which is a transversally non-degenerate critical manifold. Let us chose a small $\delta>0$. The set

$$\mathcal{W}^-_{loc}(a):=\Big\{\big(t,\theta,q,a+\partial_\theta S_a^-,\partial_q S_a^-(t,\theta,q)\big),\quad S_a^-(t,\theta,q)>-\delta\Big\}$$

is a positively invariant local stable manifold of $\mathcal{T}_1(a)$.

Similarly, S_a^+ is non-negative, it is null on $\mathbb{T}\times\mathbb{T}^m\times\{q_0\}$, and smooth around it, and this critical manifold is transversally non-degenerate. The set

$$\mathcal{W}^+_{loc}(a):=\Big\{\big(t,\theta,q,a+\partial_\theta S_a^+,\partial_q S_a^+(t,\theta,q)\big),\quad S_a^+(t,\theta,q)<\delta\Big\}$$

is a negatively invariant local unstable manifold of $\mathcal{T}_0(a)$.

The functions $S_a^\pm$ also have a global meaning. Let us give the details for S^+. For each point (T,Θ,Q), there exists a real number $\tau\in\mathbb{R}$ and at least one solution $(\theta(s),q(s)):(-\infty,\tau]\longrightarrow\mathbb{T}^m\times\mathbb{T}^r$ of the Euler-Lagrange equations such that $(\tau \bmod 1,\theta(\tau),q(\tau))=(T,\Theta,Q)$, and which is calibrated by S_a^+ in the following sense: The relation

$$\begin{aligned}&S_a^+(t \bmod 1,\theta(t),q(t))-S_a^+(s \bmod 1,\theta(s),q(s))\\&=\int_s^t L(\sigma,\theta(\sigma),q(\sigma),\dot\theta(\sigma),\dot q(\sigma))-a\dot\theta(\sigma)+a^2/2\,d\sigma\end{aligned}$$

holds for all $s<t\leqslant\tau$. The corresponding orbit is asymptotic either to $\mathcal{T}_0(a)$ or to $\mathcal{T}_1(a)$ when $s\longrightarrow-\infty$. It is not easy in general to determine whether the asymptotic torus is $\mathcal{T}_0(a)$ or $\mathcal{T}_1(a)$ but the following Lemma is not hard to prove:

Lemma 6. *If $S_a^+(T,\Theta,Q)<\xi(a)$, then each calibrated curve*

$$(\theta(s),q(s)):(-\infty,\tau]\longrightarrow\mathbb{T}^m\times\mathbb{T}^r$$

satisfying $(\tau \bmod 1,\theta(\tau),q(\tau))=(T,\Theta,Q)$, is α-asymptotic to $\mathcal{T}_0(a)$, and satisfies

$$\int_{-\infty}^{\tau} L(\sigma,\theta(\sigma),q(\sigma),\dot\theta(\sigma),\dot q(\sigma))-a\dot\theta(\sigma)+a^2/2\,d\sigma=S_a^+(T,\Theta,Q).$$

If the function S_a^+ is differentiable at (T,Θ,Q) then there is one and only one calibrated curve as above, it is characterized by the equations

$$\dot\theta(\tau)=a+\partial_\theta S_a^+(\tau,\Theta,Q),\quad \dot q(\tau)=\partial_q S_a^+(\tau,\Theta,Q).$$

Formally, the critical points of the difference $S_a^+-S_a^-$ correspond to heteroclinic orbits (in the lifted system). By studying a bit more carefully the relations between the calibrated curves and the differentiability properties of the functions

$S_a^{\pm}$ (which is one of the central aspects of Fathi's Weak KAM theory, see [29]), this idea can be made rigorous as follows:

Lemma 7. *If (T, Θ, Q) is a local minimum of the function $S_a^+ - S_a^-$, then both S_a^+ and S_a^- are differentiable at the point (T, Θ, Q), we have*

$$\left(T, \Theta, Q, a + \partial_\theta S^-, \partial_q S^-\right) = \left(T, \Theta, Q, a + \partial_\theta S^+, \partial_q S^+\right),$$

and the orbit of this point is either a heteroclinic between $\mathcal{T}_0(a)$ and $\mathcal{T}_1(a)$ or a homoclinic to $\mathcal{T}_0(a)$ or to $\mathcal{T}_1(a)$ in the system lifted to the covering, and thus it projects to an orbit homoclinic to $\mathcal{T}(a)$ in the original system.

Although it is not obvious *a priori* that a local minimum of the function $S_a^+ - S_a^-$ exists away from $q = q_0$ and $q = q_1$, this follows from the existence of minimizing heteroclinics, that we already proved. More precisely, we have:

- The minimal value of $S_a^+ - S_a^-$ is $\xi(a)$.
- The point (T, Θ, Q) is a global minimum of $S_a^+ - S_a^-$ if and only if either $Q \in \{q_0, q_1\}$ or the orbit of the point $\left(T, \Theta, Q, a + \partial_\theta S^-, \partial_q S^-\right) = \left(T, \Theta, Q, a + \partial_\theta S^+, \partial_q S^+\right)$, is a minimizing heteroclinic between $\mathcal{T}_0(a)$ and $\mathcal{T}_1(a)$.
- The set of minima of the function $S_a^+ - S_a^-$ properly contains $\mathbb{T} \times \mathbb{T}^m \times \{q_0\} \cup \mathbb{T} \times \mathbb{T}^m \times \{q_1\}$.

As a consequence, the trajectory $(\theta(t), q(t), \dot{\theta}(t), \dot{q}(t))$ is a minimizing heteroclinic if and only if $(S_a^+ - S_a^-)(t \bmod 1, \theta(t), q(t)) = \xi(a)$ for each $t \in \mathbb{R}$ (and if $q(t)$ is not identically q_0 or q_1). This minimizing heteroclinic is called isolated if, for some $t \in \mathbb{R}$, the point $(\theta(t), q(t))$ is an isolated minimum of the function

$$(\theta, q) \longmapsto (S_a^+ - S_a^-)(t \mod 1, \theta, q).$$

Now we have proved that the stable and unstable manifolds of the torus $\mathcal{T}(a)$ necessarily intersect, let us suppose that there exists a compact and connected set $A \subset \mathbb{R}^m$ such that the intersection is transversal for $a \in A$. By a continuity argument as in Proposition 1, we conclude that any two points a^- and a^+ in A can be connected by a transition chain, that is a sequence $a_0 = a^-, a_1, \ldots, a_n = a^+$ such that the unstable manifold of $\mathcal{T}(a_{i-1})$ transversally intersects the stable manifold of $\mathcal{T}(a_i)$. We would like to deduce the existence of a transversal heteroclinic orbit between $\mathcal{T}(a^-)$ and $\mathcal{T}(a^+)$, but I do not know whether the higher codimensional analog of Cresson's transitivity Lemma 4 holds. However, the weaker obstruction property proved in [21, 31] is enough to imply the existence of orbits connecting any neighborhood of $\mathcal{T}(a^-)$ to any neighborhood of $\mathcal{T}(a^+)$. It is also possible to build shadowing orbits using a variational approach. We need the slightly different assumption that $A \subset \mathbb{R}^m$ is a compact connected set such that, for all $a \in A$, all the minimizing homoclinics of $\mathcal{T}(a)$ are isolated. For each a^- and a^+ in A, it is then possible

to construct by a variational method similar to Section 2.4 a heteroclinic orbit between $\mathcal{T}(a^-)$ and $\mathcal{T}(a^+)$.

3. The General *a priori* Unstable Case

A very specific feature of all the examples studied so far is that the perturbation preserves the partially hyperbolic invariant tori $\mathcal{T}(a), a \in \mathbb{R}^m$. We now discuss the general *a priori* unstable system (2).

3.1. The Large Gap Problem. Let us assume that $r = 1$ and try to apply the method of Section 2. There is no explicit invariant torus any more, but KAM methods can be applied to prove the existence of many partially hyperbolic tori. More precisely, there exists a diffeomorphism

$$\omega_\mu(a) : \mathbb{R}^m \longrightarrow \mathbb{R}^m,$$

close to the identity, such that an invariant quasiperiodic Torus $\mathcal{T}_\mu(a)$ of frequency $\omega_\mu(a)$ exists, and is close to $\mathcal{T}(a)$, provided the frequency $\omega_\mu(a)$ satisfies some Diophantine condition. Moreover, for such values of a, the local stable and unstable manifolds $\mathcal{W}^\pm_\mu(a)$ can be generated by functions

$$S^\pm_{a,\mu}(t, \theta, q) : \mathbb{T} \times \mathbb{T}^m \times [-3/4, 3/4] \longrightarrow \mathbb{R},$$

as earlier. So we have exactly the same picture as in Section 2, except that the objects are defined only on a subset $A_\mu \subset \mathbb{R}^m$ of parameters. In order to reproduce the mechanism of Section 2, we must find elementary transitions chains $a_0, \dots, a_k$ in $\mathbb{R}^m$, with the additional requirement that $a_i \in A_\mu$. It is necessary at this point to describe a bit more the set A_μ. Roughly, the KAM methods allow to prove the existence of the Torus $\mathcal{T}_\mu(a)$ provided a belongs to

$$A_\mu = \left\{ a : \quad k \cdot (1, \omega_\mu(a)) \geqslant \frac{\sqrt{\mu}}{\|k\|^\tau} \quad \forall k \in \mathbb{Z}^{m+1} - \{0\} \right\}$$

for some constant $\tau \geqslant m+1$. This set A_μ is totally disconnected, hence it is not possible to apply a continuity method like in Proposition 1 in order to prove the existence of a transition chain. We must be more quantitative, which is possible when μ is so small that the Poincaré-Melnikov approximation is valid. In that regime, we have $\bar{\Delta}_{a,\mu} \approx \mu M_a$, where M_a has a non-degenerate critical point. The conclusion of Lemma 3 can then be proved to hold under the more explicit condition that $\|a' - a\| \leqslant C\mu$. In other words, the sequence $a_0, a_1, \dots, a_k$ is an elementary transition chain if $\|a_i - a_{i-1}\| \leqslant C\mu$. However, the gaps in A_μ have a width of size $\sqrt{\mu} > C\mu$. As a consequence, for small μ, it seems impossible to build long transition chains, and the method fails. This is the Large Gap Problem, see [35]. Even if there are classes of examples where the method can be applied because more tori exist in some regions of phase space, see [13, 19, 8] for example, the generic case seems out of range.

3.2. Normally hyperbolic invariant cylinder. The Large Gap problem has now been solved, at least in the case where $m = 1$, see [46, 17, 18, 26, 5, 44]. We will not discuss and compare all these solutions here, but just expose some general ideas which arise from them.

An important new point of view is to focus on the whole cylinder $\mathcal{C} = \cup_a \mathcal{T}(a)$ rather than on each of the tori $\mathcal{T}(a)$ individually. This cylinder is Normally hyperbolic in the sense of [34, 30], and thus it is preserved in the perturbed system. This new point of view is very natural, it appears in [43, 24], and then in many other papers. The deformed cylinder $\mathcal{C}_\mu$ contains all the preserved tori $\mathcal{T}_\mu(a)$ obtained by KAM theory. The restricted dynamics is described by an *a priori* stable system on $\mathbb{T} \times \mathbb{T}^m \times \mathbb{R}^m$. If $m > 1$, we are confronted to our lack of understanding of the *a priori* stable situation. If $m = 1$, however, the restricted system is the suspension of an area preserving twist map, and we can exploit the good understanding of these systems given by Birkhoff theory which has also been interpreted (and extended) variationally in the works of Mather [40, 41]. We consider this case ($m = 1$) from now on. The invariant 2-tori which are graphs are of particular importance (they correspond to rotational invariant circles of the time-one map). To each of these invariant graphs, we can associate two real numbers, the rotation number ω (defined from Poincaré theory of circle homeomorphisms), and the area a, which is the symplectic area of the domain of the cylinder $\mathcal{C}_\mu \cap \{t = 0\}$ delimited by the zero section and by the invariant graph under consideration. If a given invariant graph $\mathcal{T}$ of the restricted dynamics has irrational rotation number (or is completely periodic), then there is no other invariant graph with the same area a. We can take a two-covering and associate to this graph two functions $S_a^\pm$ by formula similar to (10). They generate the local stable and unstable manifold of the Torus $\mathcal{T}$, the correspond to the global minima of the difference of the so-called barrier function $S_a^+ - S_a^-$. Minimal homoclinics and isolated minimal homoclinics to $\mathcal{T}$ can be defined as in Section 2.6. The existence of minimal homoclinics can be proved basically in the same way as it was there.

Definition 8. *An invariant graph is called a transition torus if it has irrational rotation number (or if it is foliated by periodic orbits), and if all its minimal homoclinic orbits are isolated.*

Transition tori can be used to build transition chains in the same way as partially hyperbolic quasiperiodic tori with transversal homoclinics. Let $A \subset \mathbb{R}$ be the set of areas of transition tori. To each $a \in A$ is attached a unique transition torus $\mathcal{T}_\mu(a)$ (note that this torus may be only Lipschitz, and is not necessarily quasiperiodic). If A contains an interval $[a^-, a^+]$, then the existence of a heteroclinic orbit between $\mathcal{T}_\mu(a^-)$ and $\mathcal{T}_\mu(a^+)$ can be proved by already exposed methods (considering the way we have chosen our definitions, a variational method should be used, but a parallel geometric theory could certainly be given).

In general, the set A is totally disconnected, and transition chains can't be obtained by a simple continuity method. If we make the additional hypothesis

that all invariant graphs of the restricted dynamics are transition tori, then the set A is closed and a connected component $]a^-, a^+[$ of its complement corresponds to a "region of instability" of the restricted system in the terminology of Birkhoff. More precisely, the tori $\mathcal{T}_\mu(a^-)$ and $\mathcal{T}_\mu(a^+)$ enclose a cylinder which does not contain any invariant graph. The theory of Birkhoff then implies that there exist orbits of the restricted dynamics connecting an arbitrarily small neighborhood of $\mathcal{T}_\mu(a^-)$ to an arbitrarily small neighborhood of $\mathcal{T}_\mu(a^+)$. This gives an indication about how to solve the large gap problem: use the Birkhoff orbits to cross regions of instability, and the Arnold homoclinic mechanism to cross transition circles. It is by no means obvious to prove the existence of actual orbits shadowing that kind of structure. In order to do so, one should first put these mechanisms into a common framework. The variational framework seems appropriate, although a geometric approach is also possible. The Birkhoff theory was described and extended using variational methods by Mather in [40], and he proposed a new variational formalism adapted to higher dimensional situations in [41]. On the other hand, Bessi's method indicates how to put Arnold's mechanism into a variational framework. These heuristics lead to:

Theorem 3. *Let $[a^-, a^+]$ be a given interval. If all the invariant graphs of area $a \in [a^-, a^+]$ of the restricted dynamics are transition tori, then there exists an orbit $(\theta(t), q(t), \dot\theta(t), \dot q(t))$ and a time $T > 0$ such that $\dot\theta(0) \leqslant a^-$ and $\dot\theta(T) \geqslant a^+$.*

This theorem is proved using variational methods and weak KAM theory in [5], Section 11, where it is deduced from more general abstract results. It also almost follows from [18], Theorem 5.1, which is another general abstract result proved by elaborations on Mather's variational methods [41], see also [4]. Applying that result of Cheng and Yan, however, would require a minor additional generic hypothesis on the restricted dynamics. In the case where $r = 1$, a slightly weaker version of Theorem 3 could also be deduced from the earlier paper of Chen and Yan [17]. Under different sets of hypotheses, results in the same spirit have been obtained by geometric methods in [32, 33]. At the moment, these methods do not reach statements as general as Theorem 3, but they apply in contexts where the variational methods can't be used.

The following variant of Theorem 3 may deserve attention in connection to the Arnold Mechanism: Assume that a^- and a^+ belong to A, or in other words that there exist transition tori $\mathcal{T}_\mu(a^\pm)$. These tori enclose a compact invariant piece $\mathcal{C}_\mu[a^-, a^+]$ of the invariant cylinder. If all the invariant graphs contained in $\mathcal{C}_\mu[a^-, a^+]$ are transition tori, then we say that $\mathcal{C}_\mu[a^-, a^+]$ is a transition channel. The proof of Theorem 3 also implies that, if $\mathcal{C}_\mu[a^-, a^+]$ is a transition channel, then there exists a heteroclinic orbit connecting $\mathcal{T}_\mu(a^-)$ to $\mathcal{T}_\mu(a^+)$.

Theorem 3 proves the existence of diffusion under "explicit" conditions. These conditions are hard to check on a given system, but they seem to hold for typical systems. It is much harder than one may expect to prove a precise statement in that direction, but it was achieved by Cheng and Yan in [17, 18]. The main difficulty comes from the condition on the isolated minimal homoclinics. Actually, it is not hard to prove that the homoclinics to a given torus

are isolated for a typical perturbation, but we need the condition to hold for all the tori simultaneously. Since there are uncountably many tori, it is necessary to understand the regularity of the map $a \longmapsto S_a^{\pm}$. Recall that the functions $S_a^{\pm}$ are well-defined provided there exists an invariant graph of area a which has irrational rotation number or is foliated by periodic orbits. We call $\tilde{A}$ this set of areas, it contains A. Cheng and Yan prove that the map $a \longmapsto S_a^{\pm}$ is Hölder continuous on $\tilde{A}$, and deduce the genericity result using an unpublished idea of John Mather.

4. Back to the *a priori* Stable Case

The main objects in Arnold's mechanism are partially hyperbolic tori, that he called whiskered tori. It was proved by Treshchev [46], that whiskered tori exist in the *a priori* stable situation, see also [27, 45]. However, because of the Large Gap Problem, it seems difficult to prove directly the existence of transition chains made of whiskered tori. Actually, small transition chains do exist, because the density of KAM tori increases near a given one, but the length of these chains gets small when ϵ gets small, hence these chains do not produce instability of the action variables in general.

The modern paradigm on the *a priori* unstable case that we exposed in Section 3.2 elects 3-dimensional normally hyperbolic invariant cylinders as the important structure. It is well-known that normally hyperbolic invariant cylinders exist in the *a priori* stable case. For example, each 2-dimensional whiskered torus has a center manifold, which is a 3-dimensional normally hyperbolic invariant cylinder, see *e. g.*[12]. Actually, it is simpler to prove directly the existence of normally hyperbolic invariant cylinders, this involves no small divisors. However, the most direct proofs seem to produce "small" normally hyperbolic cylinders, which means that their size is getting small with ϵ, so that we face the same problem as above when we had small transition chains. The main statement of [6] is that "large" normally hyperbolic cylinders exist, meaning that their size is bounded from below independently of ϵ.

In order to be more specific, let us select a resonant momentum of the form $y_0 = (I_0, 0) \in \mathbb{R} \times \mathbb{R}^{n-1}$, with I_0 Diophantine. Assuming that the corresponding averaged potential V has a unique minimum and that this minimum is non-degenerate, we have:

Theorem 4 ([6]). *There exists two intervals $[a^-, a^+] \subset J$, J open, both* independent from ϵ, *and $\epsilon_0 > 0$ such that, for $\epsilon \in]0, \epsilon_0[$ the following holds:*

There exists a C^1 map

$$(Q, P) : \mathbb{T} \times \mathbb{T} \times J \ni (t, \theta, I) \longmapsto \big(Q(t, \theta, I), P(t, \theta, I)\big) \in \mathbb{T}^{n-1} \times \mathbb{R}^{n-1}$$

such that the flow is tangent to the graph Γ of (Q, P). Moreover, there exist two real numbers $a_0 < a^-$ and $a_1 > a^+$ in J (which depend on ϵ) such that the Treshchev tori $\mathcal{T}(a_0)$ and $\mathcal{T}(a_1)$ exist and are contained in Γ. The part Γ_0^1 of Γ

delimited by these two tori is then a compact invariant manifold with boundary of the flow, it is normally hyperbolic. It is equivalent to say that it is partially hyperbolic with a central distribution equal to the tangent space of Γ. *The inner dynamics is the suspension of an area-preserving twist map (where the area is the one induced from the ambient symplectic form).*

It is then reasonable to expect that, under generic additional hypotheses, Γ_0^1 is a transition channel as defined in Section 3.2, and thus that $\mathcal{T}(a_0)$ and $\mathcal{T}(a_1)$ are connected by a heteroclinic orbit. We are currently exploring that program in collaboration with Vadim Kaloshin and Ke Zhang. It is important to observe that the map (Q, P) is not C^1-close to $(0, 0)$, and that the inner dynamics is not close to integrable. Fortunately, Theorem 3 allows such a generality.

References

[1] V. I. Arnold, *Instability of dynamical systems with several degrees of freedom*, *Sov. Math. Doklady* **5**, 581–585 (1964).

[2] P. Bernard. *Perturbation d'un hamiltonien partiellement hyperbolique*, Comptes rendus de l'Académie des sciences. Série 1, Mathématique **323** no. 2, 189–194, (1996).

[3] P. Bernard, *Homoclinic orbits to invariant sets of quasi-integrable exact maps*, Ergodic Theory and Dynamical Systems **20** no. 6, 1583–1601 (2000).

[4] P. Bernard. *Connecting orbits of time dependent lagrangian systems,* Ann. Institut Fourier **52** no.5, 1533–1568 (2002).

[5] P. Bernard, *The dynamics of pseudographs in convex Hamiltonian systems*, J. A. M. S. **21** no. 3, 615–665 (2008).

[6] P. Bernard, *Large normally hyperbolic cylinders in a priori stable Hamiltonian systems*, preprint (2009).

[7] M. Berti, P. Bolle, *A functional analysis approach to Arnold diffusion*, Ann. I. H. P. Analyse Non Linaire **19** no. 4, 395–450 (2002).

[8] M. Berti, L. Biasco, P. Bolle, *Drift in phase space: a new variational mechanism with optimal diffusion time.* J. Math. Pures Appl. **82** no. 6, 613–664 (2003).

[9] U. Bessi, *An approach to Arnold's diffusion through the calculus of variations* Nonlinear Analysis, T. M. A., **26** no. 6, 1115–1135 (1996).

[10] U. Bessi, *Arnold's example with three rotators*, Nonlinearity, **10**, 763–781 (1997).

[11] U. Bessi, *Arnold's Diffusion with Two Resonances*, Journal of Differential Equations, **137** no. 2, 211–239 (1997).

[12] S.V. Bolotin, D.V. Treschev, *Remarks on the definition of hyperbolic tori of Hamiltonian systems* Regular and Chaotic dynamics, **5** no. 4, 401–412 (2000).

[13] S.V. Bolotin, D.V. Treschev, *Unbounded growth of energy in nonautonomous Hamiltonian systems* Nonlinearity **12**, 365–388 (1999).

[14] S.V. Bolotin, *Homoclinic orbits in invariant tori of Hamiltonian systems*, Dynamical systems in classical mechanics, 21–90, Amer. Math. Soc. Transl. Ser. 2, **168**, Amer. Math. Soc., Providence, RI, (1995).

[15] S.V. Bolotin, *Infinite number of homoclinic orbits to hyperbolic invariant tori of Hamiltonian systems*, Regular and Chaotic Dynamics **5** no. 2, 139–156 (2000).

[16] J. Bourgain, V. Kaloshin, *On diffusion in high-dimensional Hamiltonian systems,* J. Funct. Anal. **229** no. 1, 1–61 (2005).

[17] C.-Q. Cheng, J. Yan, *Existence of diffusion orbits in a priori unstable Hamiltonian systems*, J. Differential Geom. **67** no. 3, 457–517 (2004).

[18] C.-Q. Cheng, J. Yan, *Arnold Diffusion in Hamiltonian systems: the a priori unstable case*, J. Differential Geom. **82** no. 2, 229–277 (2009).

[19] L. Chierchia and G. Gallavotti, *Drift and diffusion in phase space* Ann. Inst. H. Poincaré Phys. Théor. **60** no. 1, 1–144 (1994).

[20] G. Contreras, G. Paternain, *Connecting orbits between static classes for generic Lagrangian systems*, Topology **41** no. 4, 645–666 (2002).

[21] J. Cresson, *A λ-lemma for partially hyperbolic tori and the obstruction property*, Lett. Math. Phys. **42** no. 4, 363–377 (1997).

[22] J. Cresson, *Un λ-lemme pour des tores partiellement hyperboliques*, C. R. A. S. série 1, **331** no. 1, 65–70 (2000).

[23] J. Cresson, *Temps d'instabilité des systèmes hamiltoniens initialement hyperboliques,* C. R. A. S. série 1, **332** no. 9, 831–834 (2001).

[24] A. Delshams, R. de la Llave, T. M. Seara, *A Geometric Approach to the Existence of Orbits with Unbounded Energy in Generic Periodic Perturbations by a Potential of Generic Geodesic Flows of $\mathbb{T}^2$*, Communications in Mathematical Physics, **209** no. 2, 353–392 (2000).

[25] A. Delshams, R. de la Llave, T. M. Seara, *A Geometric Mechanism for diffusion in Hamiltonian Systems Overcoming the Large Gap Problem: Heuristics and Rigorous Verification on a Model,* Mem. A.M.S. **179** no. 844 (2006).

[26] A. Delshams, R. de la Llave, T. M. Seara, Orbits of unbounded energy in quasi-periodic perturbations of geodesic flows, *Adv. in Math.* **202**, 64–188 (2006).

[27] L. H. Eliasson, *Biasymptotic solutions of perturbed integrable Hamiltonian systems*, Bull. Braz. Math. Soc. **25** no. 1, 57–76 (1994).

[28] A. Fathi, Orbites heteroclines et ensemble de Peierls, *Comptes rendus de l'Académie des sciences. Série 1, Mathématique* **326** no. 10, 1213–1216 (1998).

[29] A. Fathi, *Weak KAM theorem in Lagrangian dynamics.* ghost book.

[30] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J. **21**, 193–226 (1971).

[31] E. Fontich, P. Martin, *Differentiable invariant manifolds for partially hyperbolic tori*, Nonlinearity **13**, 1561–1593 (2000).

[32] M. Gidea, C. Robinson, *Shadowing orbits for transition chains of invariant tori*, Nonlinearity **20**, 1115–1143, (2007).

[33] M. Gidea, C. Robinson, *Obstruction argument for transition chains of Tori interspersed with gaps* Discrete Contin. Dyn. Syst. Ser. S **2** no. 2, 393–416 (2009).

[34] M.W. Hirsch, C.C. Pugh, M. Shub, *Invariant manifolds*, Lecture notes in Math. Springer Berlin, New York, (1977).

[35] P. Lochak, *Arnold diffusion; a compendium of remarks and questions*, in *Hamiltonian systems with three or more degrees of freedom*, 168–213, Kluwer Academic Publishers, (1999).

[36] P. Lochak, J-P. Marco, *Diffusion times and stability exponents for nearly integrable analytic systems,* Cent. Eur. J. Math. **3** no. 3, 342–397 (2005).

[37] P. Lochack, J. P. Marco, D. Sauzin, *On the Splitting of Invariant Manifolds in Multidimensional Near-Integrable Hamiltonian Systems*, Mem. A.M.S. **163** no. 775 (2003).

[38] J. P. Marco, *Transition le long des chaines de tores invariants pour les systèmes hamiltoniens analytiques*, Annales de l'I. H. P. Physique théorique **64** no. 2, 205–252 (1996).

[39] J. P. Marco, D. Sauzin, *Stability and instability for Gevrey quasi-convex near-integrable Hamiltonian systems*, Publ. Math. Inst. Hautes tudes Sci. **96**, 199–275 (2003).

[40] J. N. Mather, *Variational construction of orbits of twist diffeomorphisms*, J.A.M.S. **4** no. 2, 207–263 (1991).

[41] J. N. Mather, *Variational construction of connecting orbits*, Ann. Inst. Fourier **43**, 1349–1368 (1993).

[42] J. N. Mather, *Arnold diffusion: announcement of results*, J. Math. Sci. (N. Y.) **124** no. 5, 5275–5289 (2004).

[43] R. Moeckel, *Transition Tori in the Five-Body Problem*, J.D.E. **129**, 290–314 (1996).

[44] R. Moeckel, *Generic drift on Cantor sets of annuli*, in *Celestial Mechanics*, Contemp. Math. 292, A.M.S., 163–171 (2002).

[45] L. Niederman, *Dynamics around simple resonant tori in nearly integrable Hamiltonian systems,* J. Differential Equations **161** no. 1, 1–41 (2000).

[46] D. Treshchev, *The Mechanism of destruction of resonance tori of Hamiltonian systems*, Math. USSR Sb. **68** no. 1, 181–203 (1991).

[47] D. Treshchev, *Evolution of slow variables in a priori unstable Hamiltonian systems*, *Nonlinearity* **17** no. 5, 1803–1841 (2004).

[48] Z. Xia, *Arnold diffusion and instabilities in Hamiltonian dynamics*, preprint (2002).

[49] K. Zhang, *Speed of Arnold diffusion for analytic Hamiltonian systems*, preprint (2009).

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Quadratic Julia Sets with Positive Area

Xavier Buff* and Arnaud Chéritat†

Abstract

We recently proved the existence of quadratic polynomials having a Julia set with positive Lebesgue measure. We present the ideas of the proof and the techniques involved.

Mathematics Subject Classification (2010). Primary 37F50; Secondary 37F25.

Keywords. Holomorphic dynamics, Julia sets, small divisors.

1. Introduction

We study the dynamics of polynomials $P : \mathbb{C} \to \mathbb{C}$, i.e., the sequences defined by induction:

$$z_0 \in \mathbb{C}, \quad z_{n+1} = P(z_n).$$

The sequence (z_n) is called the *orbit* of z_0.

Definition 1. The filled-in Julia set $K(P)$ is the set of points $z_0 \in \mathbb{C}$ with bounded orbits. The Julia set $J(P)$ is the boundary of $K(P)$.

The filled-in Julia set $K(P)$ is a compact subset of $\mathbb{C}$ and so, its boundary $J(P)$ has empty interior. Points outside $K(P)$ have an orbit tending to ∞.

This subject has its roots in complex analysis, strongly linked to Montel's theorem on normal families. In particular, the family of iterates $(P^{\circ n})_{n\geq 0}$ is normal on the complement of $J(P)$ (called the Fatou set of P) and on any open set intersecting the Julia set $J(P)$, the sequence of iterates is not normal, since such an open set contains points with bounded orbit and points whose orbit tends to ∞. Thus, the Julia set $J(P)$ may be viewed as the *chaotic set* for the dynamics of P.

Periodic points play an important role from a dynamical point of view. A *periodic point* of P of period p is a point z such that $P^{\circ p}(z) = z$ for some

*Université de Toulouse; UPS, INSA, UT1, UTM; Institut de Mathématiques de Toulouse; F-31062 Toulouse, France. E-mail: xavier.buff@math.univ-toulouse.fr.

†CNRS; Institut de Mathématiques de Toulouse UMR 5219; F-31062 Toulouse, France. E-mail: arnaud.cheritat@math.univ-toulouse.fr.

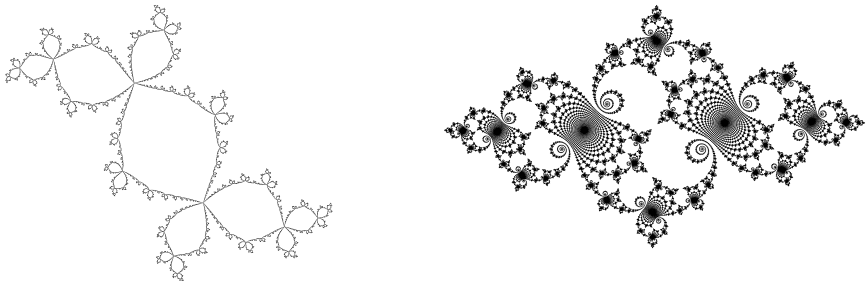


Figure 1. Left: the Julia set of a quadratic polynomial for which the critical point is periodic of period 3. It is known as the Douady Rabbit. Right: the Julia set of a quadratic polynomial with an unbounded critical orbit. The Julia set is a Cantor set.

smallest integer $p \geq 1$. The set $\{z, P(z), \dots, P^{\circ(p-1)}(z)\}$ is a *periodic cycle*. The periodic point is repelling (respectively attracting, superattracting, indifferent) if its *multiplier* $\lambda = (P^{\circ p})'(z)$ satisfies $|\lambda| > 1$ (respectively $0 < |\lambda| < 1$, $\lambda = 0$, $|\lambda| = 1$). The Julia set $J(P)$ may equivalently be defined as the closure of the set of repelling periodic points of P.

Fatou observed that the dynamics of a polynomial P is intimately related to the behavior of the orbit of the critical points of P. A *critical point* of P is a point $\omega \in \mathbb{C}$ for which $P'(\omega) = 0$. In particular, Fatou proved that $K(P)$ is connected if and only if all the critical points of P are in $K(P)$. Further, when all the critical points of P are in the complement of $K(P)$, then $K(P) = J(P)$ is a Cantor set.

Fatou suggested that one should apply to $J(P)$ the methods of Borel-Lebesgue for the measure of sets. This naturally yields the following question.

Question. *What can we say about the Lebesgue measure of the Julia set of a polynomial?*

Until recently, the common belief was that Julia sets of polynomials always had area (Lebesgue measure) zero. It is known that the area of $J(P)$ is zero in several cases, in particular when $J(P)$ does not contain critical points of P or when the orbit of any critical point of P contained in $J(P)$ is finite ([DH] or [L1]).

In the rest of the article, we will mainly focus on the case of quadratic polynomials

$$Q_\lambda(z) = \lambda z + z^2 \quad \text{with} \quad \lambda \in \mathbb{C}.$$

Such a polynomial has a fixed point at 0 with multiplier λ and a unique critical point $\omega_\lambda = -\lambda/2$. So, we have the following dichotomy: $K(Q_\lambda)$ is connected if the orbit of ω_λ is bounded and is a Cantor set otherwise. We shall denote by $\mathcal{M}$ the set of parameters $\lambda \in \mathbb{C}$ for which $K(Q_\lambda)$ is connected (see Figure 2).

The area of $J(Q_\lambda)$ is zero:

- when λ is outside the connectivity locus $\mathcal{M}$;

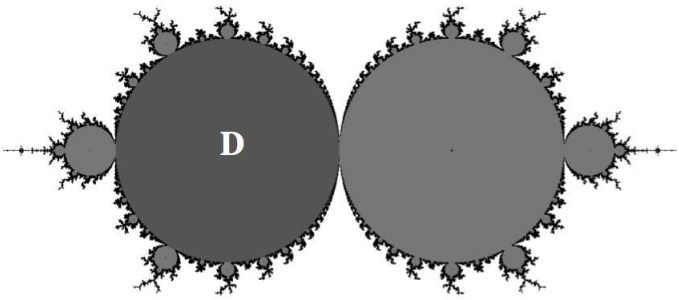


Figure 2. The set $\mathcal{M}$ of parameters $\lambda \in \mathbb{C}$ for which $J(Q_\lambda)$ is connected. It contains the unit disk $\mathbb{D}$ for which Q_λ has an attracting fixed point at 0.

- when Q_λ has a (super)attracting cycle (conjecturally, this is true for all λ in the interior of $\mathcal{M}$, and according to [MSS], if there were a parameter λ in the interior of $\mathcal{M}$ for which Q_λ does not have an attracting cycle, it is known that $J(Q_\lambda)$ would necessarily have positive area);
- for a generic (in the sense of Baire) λ in the boundary of $\mathcal{M}$ ([L1] or [L2]),
- if Q_λ is not infinitely renormalizable ([L3] or [Sh]), a condition that we will not define here;
- if $\lambda = e^{2i\pi\alpha}$ with $\alpha = \mathrm{a}_0 + \cfrac{1}{\mathrm{a}_1 + \cfrac{1}{\mathrm{a}_2 + \ddots}}$ and $\log \mathrm{a}_n = \mathcal{O}(\sqrt{n})$ ([PZ]); this condition on α holds for almost every $\alpha \in \mathbb{R}/\mathbb{Z}$.

In the 1990's, Douady proposed a program to show that there exist complex numbers λ of modulus 1 so that the area of $J(Q_\lambda)$ is positive. After a major breakthrough by the second author [C1], we finally brought Douady's program to completion in 2005. For a presentation of Douady's initial program, the reader is invited to consult [C2].

Theorem 1.1 ([BC2]). *There exist λ of modulus 1 such that $J(Q_\lambda)$ has positive area.*

We will present the ideas of the proof and the techniques involved.

2. Quadratic Polynomials with an Indifferent Fixed Point

We may classify the quadratic polynomials Q_λ with $|\lambda| = 1$ in three categories as follows. First, let us note $\lambda = e^{2\pi i\alpha}$ with $\alpha \in \mathbb{R}/\mathbb{Z}$ and set

$$P_\alpha(z) = e^{i2\pi\alpha}z + z^2, \quad K_\alpha = K(P_\alpha) \quad \text{and} \quad J_\alpha = J(P_\alpha).$$

If $\alpha \in \mathbb{Q}/\mathbb{Z}$, we say that 0 is a parabolic fixed point of P_α. In that case, K_α has interior and $0 \in J_\alpha$. The orbit of a point in the interior of K_α converges to 0. The Julia set J_α has area zero.

If $\alpha \in (\mathbb{R} - \mathbb{Q})/\mathbb{Z}$, the dynamical behavior of P_α near 0 depends subtly on the arithmetical properties of α. We have the following dichotomy.

- If α is sufficiently Liouville, then $J_\alpha = K_\alpha$. Any neighborhood of 0 contains points with bounded orbit and points whose orbit tends to ∞. Cremer proved that the set of such angles α is G_δ dense in $\mathbb{R}/\mathbb{Z}$. We say that P_α has a Cremer fixed point at 0.

- If α is badly approximated by rational numbers, then 0 is in the interior of K_α. In that case, we denote by Δ_α the component of the interior of K_α that contains 0. Then P_α is holomorphically conjugate to the aperiodic rotation $R_\alpha : z \mapsto e^{2\pi i\alpha} z$: there is an analytic isomorphism ϕ between the unit disk $\mathbb{D}$ and Δ_α such that $\phi(0) = 0$ and $\phi \circ R_\alpha = P_\alpha \circ \phi$. One says that the polynomial P_α is *linearizable* and the component Δ_α is called a *Siegel disk*. Siegel [Si] proved that this property holds when α is Diophantine, in particular for a set of full measure in $\mathbb{R}/\mathbb{Z}$ (α is Diophantine if there are constants $c > 0$ and $\tau \geq 2$ such that $|\alpha - p/q| > c/q^\tau$ for all rational numbers p/q).

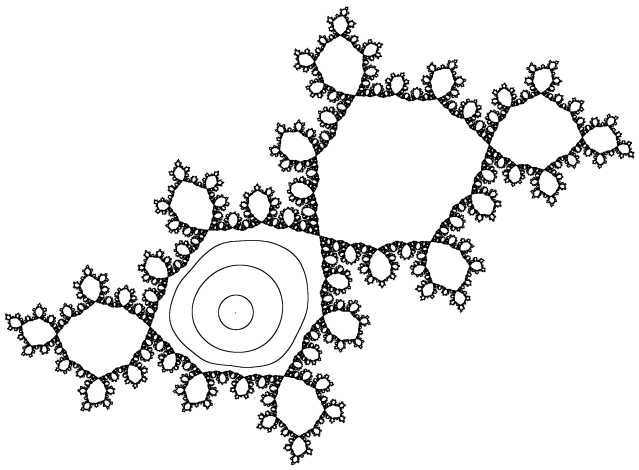


Figure 3. The Julia set of P_α for $\alpha = (\sqrt{5} - 1)/2$. We have drawn the orbits of some points in the Siegel disk. Each orbit accumulates on a $\mathbb{R}$-analytic circle.

In fact, there is a complete arithmetic characterization of the two previous sets of angles. Let $(p_n/q_n)_{n\geq 0}$ be the approximants to α given by the continued fraction algorithm. Brjuno [Brj] proved that when

$$B(\alpha) = \sum_{n\geq 0} \frac{\log q_{n+1}}{q_n} < +\infty,$$

the polynomial P_α is linearizable. Yoccoz [Y] proved that when $B(\alpha) = +\infty$, the polynomial P_α has a Cremer fixed point at 0. In addition, any neighborhood of 0 contains a cycle which is not reduced to $\{0\}$.

We have the following refined versions of our theorem.

Theorem 2.1. *There exist angles $\alpha \in (\mathbb{R} - \mathbb{Q})/\mathbb{Z}$ for which P_α has a Cremer fixed point at 0 and* $\text{area}(J_\alpha) > 0$.

Theorem 2.2. *There exist angles $\alpha \in (\mathbb{R} - \mathbb{Q})/\mathbb{Z}$ for which P_α has a Siegel disk and* $\text{area}(J_\alpha) > 0$.

We will now sketch the proof of the first theorem. The proof of the second theorem relies on similar ideas.

3. Strategy of the Proof

Proposition 3.1. *The function $\alpha \mapsto \text{area}(K_\alpha) \in [0, +\infty[$ is upper semi-continuous.*

In other words, if $\alpha_n \to \alpha$, then

$$\limsup_{n\to\infty} \text{area}(K_{\alpha_n}) \leq \text{area}(K_\alpha).$$

Proof. Every open set containing K_α contains $K_{\alpha'}$ for α' close enough to α. □

We shall see that the existence of Julia sets with positive area is an immediate consequence of the following key proposition which is illustrated by Figure 4.

Proposition 3.2. *There exists a non empty set $\mathcal{S}$ of Diophantine numbers such that: for all $\alpha \in \mathcal{S}$ and all $\varepsilon > 0$, there exists $\alpha' \in \mathcal{S}$ with*

- $|\alpha' - \alpha| < \varepsilon$,
- *$P_{\alpha'}$ has a cycle in $D(0, \varepsilon) \setminus \{0\}$ and*
- $\text{area}(K_{\alpha'}) \geq (1 - \varepsilon)\text{area}(K_\alpha)$.

With this proposition, one concludes as follows. First, we choose ε_n in $(0, 1)$ so that $\prod(1 - \varepsilon_n) > 0$. Then, we construct $(\theta_n \in \mathcal{S})$ so that:

- (θ_n) is a Cauchy sequence.
- $\text{area}(K_{\theta_n}) \geq (1 - \varepsilon_n)\text{area}(K_{\theta_{n-1}})$.
- For $\theta = \lim \theta_n$, the polynomial P_θ has small cycles.

Since θ_n is Diophantine, $K(P_{\theta_n})$ has non empty interior and so, its area is positive. Since P_θ has small cycles, it is not linearizable, and so $J_\theta = K_\theta$. By upper semi-continuity of the function $\alpha \mapsto \text{area}(K_\alpha)$, we have

$$\text{area}(J_\theta) = \text{area}(K_\theta) \geq \limsup_{n\to+\infty} \text{area}(K_{\theta_n}) \geq \text{area}(K_{\theta_0}) \cdot \prod_{n\geq 1} (1 - \varepsilon_n) > 0.$$

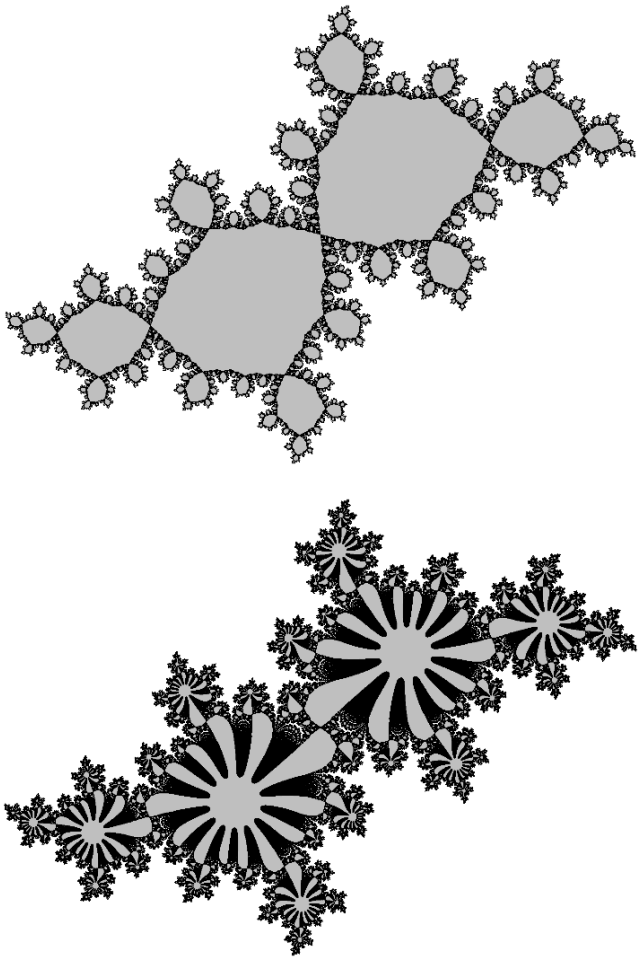


Figure 4. Two filled-in Julia sets K_α (top) and $K_{\alpha'}$ (bottom), with α' a well-chosen perturbation of α. If α and α' are chosen carefully enough the loss of measure from K_α to $K_{\alpha'}$ is small.

4. The Set $\mathcal{S}$

For $\alpha \in \mathbb{R} - \mathbb{Q}$, let us use the continued fraction notation

$$[a_0, a_1, a_2, \ldots] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}}.$$

Recall that an irrational number $\alpha = [a_0, a_1, a_2, \ldots]$ is of *bounded type* if the sequence (a_n) is bounded.

Definition 2. For $N \geq 1$, denote by $\mathcal{S}_N$ be the set of irrational numbers $\alpha = [\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_2, \ldots]$ of bounded type such that $\mathrm{a}_n \geq N$ for all $n \geq 1$.

Proposition 4.1. *If $N \geq 1$ is a sufficiently large integer, then Proposition 3.2 holds with $\mathcal{S} = \mathcal{S}_N$.*

5. McMullen's Results on Siegel Disks of Bounded Type

As we shall see below, the proof of Proposition 4.1 reduces to the following result that is illustrated on Figure 5.

Lemma 5.1. *If $N \geq 1$ is a sufficiently large integer, then for all $\alpha \in \mathcal{S}_N$, there is a sequence $\alpha_n \in \mathcal{S}_N$ converging to α such that*

- *P_{α_n} has a cycle converging to 0 as $n \to \infty$,*
- *for all open set $U \subset \Delta_\alpha$, we have*

$$\liminf_{n\to\infty} \mathrm{area}(U \cap \Delta_{\alpha_n}) \geq \frac{1}{2}\mathrm{area}(U \cap \Delta_\alpha) \quad \textit{and}$$

- *$\overline{\Delta}_{\alpha_n} \to \overline{\Delta}_\alpha$ for the Hausdorff topology on compact subsets of $\mathbb{C}$.*

The second assertion says that asymptotically, the Siegel disks Δ_{α_n} are at least 1/2-dense in Δ_α.

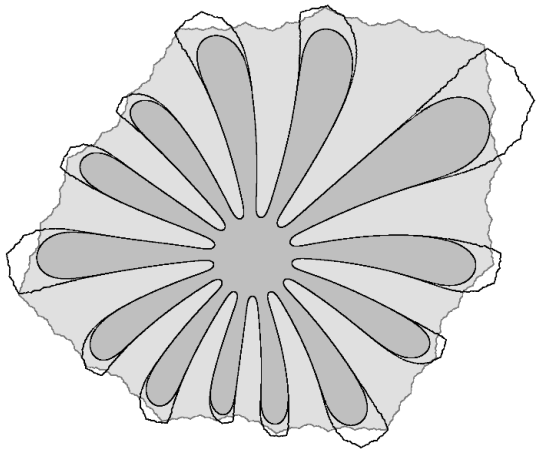


Figure 5. Illustration of lemma 5.1 for $N = 1$, $\alpha = [0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 \ldots]$ and $\alpha' = [0, 1, 1, 1, 1, 1, 1, 1, 1, 10^{10}, 1, 1, \ldots]$. The Siegel disk Δ_α is colored light grey. The boundary of the Siegel disk $\Delta_{\alpha'}$ is drawn. The darker set is the set of points in $\Delta_{\alpha'}$ whose orbit under iteration of $P_{\alpha'}$ remains in Δ_α.

We then use an argument of *toll belts* inspired by work of McMullen [McM] to promote the loss of $1/2$ for the area of Siegel disks to an arbitrarily small loss for the area of the filled-in Julia sets. For the argument of toll belts to work, we need that α is of bounded type and $\overline{\Delta}_{\alpha_n} \to \overline{\Delta}_\alpha$ as $n \to \infty$. More precisely, we use the following result of McMullen.

Theorem 5.2 (McMullen). *Assume α is a bounded type irrational and $\delta > 0$. Then, every point $z \in \partial\Delta_\alpha$ is a Lebesgue density point of the set $K(\delta)$ of points whose orbit under iteration of P_α remains at distance less than δ from Δ_α and eventually intersect Δ_α.*

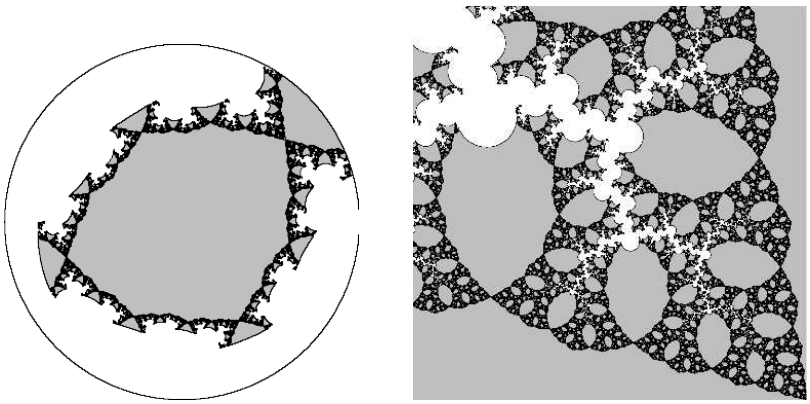


Figure 6. If $\alpha = (\sqrt{5}-1)/2$, the critical point of P_α is a Lebesgue density point of the set of points whose orbit remain in $D(0,1)$. Left: the set of points whose orbit remains in $D(0,1)$. Right: a zoom near the critical point.

Proof of Proposition 4.1 assuming lemma 5.1. Assume $\alpha \in \mathcal{S}_N$ and let $(\alpha_n)_{n\geq 0}$ be a sequence of $\mathcal{S}_N$ given by lemma 5.1. Denote by K (resp. K_n) the filled-in Julia set of P_α (resp. P_{α_n}) and by Δ (resp. Δ_n) its Siegel disk. We know that asymptotically, the Siegel disks Δ_n are at least $1/2$-dense in the Siegel disk Δ. We want to show that $\text{area}(K_n) \to \text{area}(K)$, which amounts to proving that the density of K_n in Δ tends to 1 as $n \to \infty$.

For all S, one can find a finite nested sequence of *toll belts* $(W_s)_{1\leq s\leq S}$

$$W_s = \left\{ z \in \mathbb{C} \ ; \ 2\delta_s < d(z,\Delta) < 8\delta_s \right\} \quad \text{with} \quad 8\delta_{s+1} < \delta_s,$$

surrounding the Siegel disk Δ such that for n large enough the following holds.

- The orbit under iteration of P_{α_n} of any point in $\Delta \setminus K_n$ must pass through all the toll belts.
- Thanks to Lemma 5.1, the toll belts surround the Siegel disk Δ_n.

- Thanks to Theorem 5.2 and Lemma 5.1, under the iterates of P_{α_n}, at least $1/2-\varepsilon$ of points in the toll belt W_{s+1} will be captured by the Siegel disk Δ_n without being able to enter the toll belt W_s.
- Since the toll belts surround the Siegel disk Δ_n, they are free of the postcritical set of P_{α_n}. This gives us Koebe control of points passing through the belt, implying that at most $1/2+\varepsilon$ of points in Δ that manage to reach W_{s+1} under iteration of P_{α_n} will manage to reach W_s.

As a consequence, at most $(1/2+\varepsilon)^S$ points in Δ can have an orbit under iteration of P_{α_n} that passes through all the belts and we are done by choosing S large enough. □

6. The Sequence (α_n)

We claim that if N is a large enough integer and if $\alpha=[\mathrm{a}_0,\mathrm{a}_1,\ldots]\in\mathcal{S}_N$, then Lemma 5.1 holds for the sequence (α_n) defined by

$$\alpha_n=[\mathrm{a}_0,\mathrm{a}_1,\ldots,\mathrm{a}_n,A_n,N,N,N,\ldots] \tag{1}$$

where the sequence (A_n) is chosen so that

$$A_n\geq N,\quad \sqrt[q_n]{A_n}\underset{n\to+\infty}{\longrightarrow}+\infty\quad\text{and}\quad \sqrt[q_n]{\log A_n}\underset{n\to+\infty}{\longrightarrow}1.$$

Lemma 5.1 has three parts which can be treated one at a time: the existence of a cycle of P_{α_n} close to 0, the density of the perturbed Siegel disk Δ_{α_n} within Δ_α and the Hausdorff convergence of $\overline{\Delta}_{\alpha_n}$ to $\overline{\Delta}_\alpha$.

7. The Control of the Cycle

In order to prove the existence of a cycle of P_{α_n} close to 0, we use a result of the second author [C1].

Proposition 7.1. *Assume P_α has a Siegel disk. Let (p_n/q_n) be the approximant to α given by the continued fraction algorithm. Let $\chi:\mathbb{D}\to\Delta_\alpha$ be an isomorphism which sends 0 to 0. There is a sequence (r_n) converging to 1 and a sequence of univalent maps $(\chi_n:D(0,r_n)\to\Delta_\alpha)$ converging locally uniformly to $\chi:\mathbb{D}\to\Delta_\alpha$ such that the following holds: if (α_n) is a sequence converging to α with $\limsup\sqrt[q_n]{|\alpha_n-p_n/q_n|}<1$ and if C_n is the set of q_n-th roots of α_n-p_n/q_n, then for n large enough, $\chi_n(C_n)$ is a cycle of period q_n of P_{α_n}.*

The functions $\chi_n:D(0,r_n)\to\Delta_\alpha$ are called *explosion functions*. They control the explosion, as α goes away from p_n/q_n, of the cycle of period q_n of P_α which coalesces at 0 when $\alpha=p_n/q_n$.

Now, observe that the sequence (α_n) defined by Equation (1) satisfies

$$\alpha_n - \frac{p_n}{q_n} \underset{n\to\infty}{\sim} \frac{(-1)^n}{q_n^2 A_n}.$$

Since $\sqrt[q_n]{A_n} \to +\infty$, we have that $\sqrt[q_n]{|\alpha_n - p_n/q_n|} \to 0$.

Thus, for n sufficiently large, the set C_n of q_n-th roots of $\alpha_n - p_n/q_n$ is contained in an arbitrarily small neighborhood of 0. The sequence (χ_n) converges locally uniformly to χ. So, for n large enough, the set $\chi_n(C_n)$, which is a cycle of P_{α_n}, is contained in an arbitrarily small neighborhood of 0

8. The Density of Perturbed Siegel Disks

We still assume that (α_n) is defined by Equation (1).

In order to control the density of the Siegel disks Δ_{α_n} within the Siegel disk Δ_α, we may work in the coordinates given by the explosion functions χ_n. In other words, we set

$$f_n = \chi_n^{-1} \circ P_{\alpha_n} \circ \chi_n.$$

As $n \to \infty$, the domain of f_n eventually contains any compact subset of $\mathbb{D}$. The sequence (f_n) converges locally uniformly to the rotation R_α. The map f_n fixes 0 with derivative $e^{2\pi i\alpha_n}$ and has a Siegel disk Δ_n whose image by χ_n is contained in the Siegel disk Δ_{α_n} of P_{α_n}.

We want to prove that asymptotically as $n \to \infty$, the Siegel disks Δ_n are 1/2-dense in the unit disk. For this purpose, it is not enough to compare the dynamics of f_n with the dynamics of a rotation. Instead, we will compare it with the (real) dynamics of an appropriate polynomial vector field ξ_n.

Note that by property of the explosion functions χ_n, the set C_n of q_n-th roots of $\varepsilon_n = \alpha_n - p_n/q_n$ is a periodic cycle of f_n of period q_n. Let ξ_n be the polynomial vector field which has simple roots exactly at 0 and the points of C_n and which has derivative $2\pi i q_n \varepsilon_n$ at 0. Then, the time-1 map of ξ_n fixes 0 and the points of C_n (which are also fixed points of $f_n^{\circ q_n}$) with multiplier $e^{2\pi i q_n \varepsilon_n}$ at 0 (which is also the multiplier of $f_n^{\circ q_n}$ at 0). Thanks to those properties, there is a good hope that the time-1 map of ξ_n very well approximates $f_n^{\circ q_n}$. This vector field is

$$\xi_n = \xi_n(z)\frac{\mathrm{d}}{\mathrm{d}z} = 2\pi i q_n z(\varepsilon_n - z^{q_n})\frac{\mathrm{d}}{\mathrm{d}z}.$$

We have an explicit description of the vector field ξ_n which is invariant under the rotation $z \mapsto e^{2\pi i/q_n} z$. For all $\rho < 1$ and all n sufficiently large, the set $X_n(\rho)$ defined below is invariant under the real flow of the vector field ξ_n:

$$X_n(\rho) = \left\{ z \in \mathbb{C} \ ; \ \frac{z^{q_n}}{z^{q_n} - \varepsilon_n} \in D(0, s_n) \right\} \quad \text{with} \quad s_n = \frac{\rho^{q_n}}{\rho^{q_n} + |\varepsilon_n|}.$$

This set looks like an amoeba with q_n arms. Asymptotically, the density of $X_n(\rho)$ in $D(0,\rho)$ is at least 1/2.

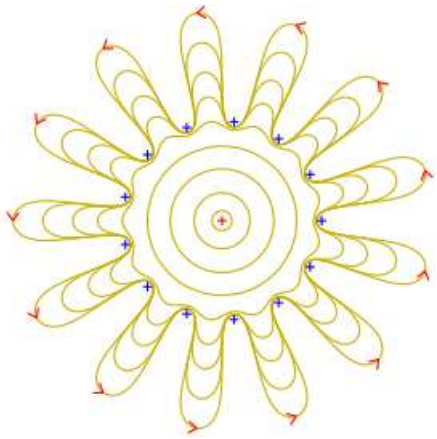


Figure 7. Some real trajectories for the vector field ξ_n; zeroes of the vector field are shown.

Using very careful estimates on how close $f_n^{\circ q_n}$ is to the time-1 map of the vector field ξ_n and using Yoccoz renormalization techniques [Y], we obtain the following result which implies the required control on the asymptotic density of Δ_n within $\mathbb{D}$.

Proposition 8.1. *For all $\rho < 1$, if n is large enough, the set $X_n(\rho)$ is contained in the Siegel disk Δ_n of f_n.*

9. Hausdorff Convergence of Perturbed Siegel Disks

In order to prove the Hausdorff convergence of $\overline{\Delta}_{\alpha_n}$ to $\overline{\Delta}_\alpha$, we use techniques of *near parabolic renormalization* introduced recently by Inou and Shishikura [IS]. Those techniques are far too elaborate for us to present them here.

Let us however insist on the fact that it is to apply those techniques that we have to assume that the entries in the continued fraction expansion of α are large enough ($\mathrm{a}_n \geq N$ for all n).

10. Further Questions

Our proof of existence of quadratic polynomials having a Julia set of positive area is *a priori* not constructive. It would be interesting to have informations regarding the set of $\alpha \in \mathbb{R}$ for which the Julia set J_α has positive area.

Theorem 10.1 (Petersen, Zakeri). *For almost every $\alpha \in \mathbb{R}$, we have* $\mathrm{area}(J_\alpha) = 0$.

Question. *Is the set of parameters $\alpha \in \mathbb{R}$ for which* $\operatorname{area}(J_\alpha) > 0$ *a G_δ-dense set?*

Now that we have proved the existence of $\alpha \in \mathbb{R} - \mathbb{Z}$ for which $J_\alpha = K_\alpha$ has positive area, we can change the question. Indeed, we do not know of a single example of a non linearizable quadratic polynomial P_α with $\alpha \in \mathbb{R} - \mathbb{Q}$ for which the Julia set has area zero. It may well be that all such Julia set have positive area.

Question. *Is there $\alpha \in \mathbb{R}$ such that $J_\alpha = K_\alpha$ and* $\operatorname{area}(J_\alpha) = 0$*?*

A key point in our proof was the observation that the function $\alpha \mapsto \operatorname{area}(K_\alpha)$ is upper semicontinuous. It would be interesting to have additional informations regarding its continuity properties.

Theorem 10.2 (Douady). *The function $\alpha \mapsto \operatorname{area}(K_\alpha)$ is discontinuous at rational numbers.*

Question. *Is the function $\alpha \mapsto \operatorname{area}(K_\alpha)$ continuous at irrational numbers?*

The techniques we have been developing for studying the area of Julia sets already had fruitful applications, in particular for the study of Siegel disks. Answering a question of Herman, we proved the following result in collaboration with A. Avila.

Theorem 10.3 ([ABC]). *There exist $\alpha \in \mathbb{R}$ such that P_α has a Siegel disk whose boundary is a smooth (C^∞) Jordan curve.*

In that case, the boundary of the Siegel disk P_α cannot contain the critical point of P_α. This is in contrast to the following result of Petersen and Zakeri.

Theorem 10.4 (Petersen-Zakeri). *For almost every $\alpha \in \mathbb{R}$, P_α has a Siegel disk whose boundary is a Jordan curve passing through the critical point $\omega_\lambda = -\lambda/2$.*

This raises naturally the following questions.

Question. *If P_α has a Siegel disk, is the boundary of Δ_α always a Jordan curve?*

Question. *For which values of α does P_α have a Siegel disk whose boundary contains the critical point?*

References

[ABC] A. Avila, X. Buff, A. Chéritat, *Siegel disks with smooth boundaries*, Acta Mathematica (2004) 193, 1–30.

[Brj] A.D. Brjuno, *Analytic forms of differential equations*, Trans. Mosc. Math. Soc. **25** (1971).

[BC1] X. Buff & A. Chéritat, *Upper Bound for the Size of Quadratic Siegel Disks*, Invent. Math. (2004) 156/1, 1–24.

[BC2] X. Buff & A. Chéritat, *Ensembles de Julia quadratiques de mesure de Lebesgue strictement positive*, Comptes Rendus Mathématiques (2005) 341/11, 669–674.

[C1] A. Chéritat, *Recherche d'ensembles de Julia de mesure de Lebesgue positive*, Thèse, Orsay (2001).

[C2] A. Chéritat, *The hunt for Julia sets with positive measure*, Complex dynamics, 539–559, A K Peters, Wellesley, MA, 2009.

[DH] A. Douady & J.H. Hubbard *Etude dynamique des polynômes complexes* I & II, Publ. Math. d'Orsay (1984–85).

[IS] H. Inou & M. Shishikura, *The renormalization for parabolic fixed points and their perturbation*, Preprint.

[L1] M. Lyubich, *On the typical behavior of the trajectories of a rational mapping of the sphere*, Dokl. Acad. Nauk SSSR **68** (1982), 29–32 (translated in Soviet Math. Dokl. **27** (1983) 22–25).

[L2] M. Lyubich, *An analysis of the stability of the dynamics of rational functions*, Teor. Funktsii, Funk. Analiz i Pril **42** (1984) 72–91 (translated on Selecta Mathematics Sovetica **9:1** (1990) 69–90.

[L3] M. Lyubich, *On the Lebesgue measure of the Julia set of a quadratic polynomial*, Stonybrook IMS Preprint 1991/10.

[McM] C.T. McMullen, *Self-similarity of Siegel disks and Hausdorff dimension of Julia sets*, Acta Math., **180** (1998), 247–292.

[MSS] R. Mañé, P. Sad & D.P. Sullivan, *On the dynamics of rational maps*, Ann. Sci. Éc. Norm. Sup., Paris, 16:193–217, (1983).

[PZ] C.L. Petersen & S. Zakeri, *On the Julia set of a typical quadratic polynomial with a Siegel disk*, Ann. of Math. **159** (2004) 1–52.

[Sh] M. Shishikura, *Topological, geometric and complex analytic properties of Julia sets*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), 886–895, Birkhäuser, Basel, 1995.

[Si] C.L. Siegel, *Iteration of analytic functions*, Ann. of Math. **vol** 43 (1942).

[Y] J.C. Yoccoz, *Petits diviseurs en dimension 1*, S.M.F., Astérisque **231** (1995).

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Variational Construction of Diffusion Orbits for Positive Definite Lagrangians

Chong-Qing Cheng*

Abstract

In this lecture, we sketch the variational construction of diffusion orbits in positive definite Lagrangian systems. Diffusion orbits constructed this way connects different Aubry sets, along which the action is locally minimized.

Mathematics Subject Classification (2010). Primary 37Jxx; Secondary 70Hxx.

Keywords. Tonelli Lagrangian, Action minimizing, Arnold diffusion.

1. Introduction

The variational method we are discussing here is based on Mather's theory for Tonelli Lagrangian systems. Let M be an n-dimensional smooth manifold without boundary. A function $L : TM \times \mathbb{T}$ is called Tonelli Lagrangian if it satisfies the following conditions:

Positive definiteness. For each $(x, t) \in M \times \mathbb{T}$, the Lagrangian function is strictly convex in velocity: the Hessian $L_{\dot{x}\dot{x}}$ is positive definite.

Super-linear growth. We assume that L has fiber-wise superlinear growth: for each $(x, t) \in M \times \mathbb{T}$, we have $L/\|\dot{x}\| \to \infty$ as $\|\dot{x}\| \to \infty$.

Completeness. All solutions of the Lagrangian equation are well defined for all $t \in \mathbb{R}$.

Given a cohomology class c, let η_c be a closed 1-form such that its cohomology class $[\eta_c] = c$. Denote by $-\alpha(c)$ the c-minimal average action. A curve $\gamma : \mathbb{R} \to M$ is called c-minimal (c-semi static) if

$$\int_{t_0}^{t_1} (L - \eta_c + \alpha(c))(d\gamma(s), s)ds \leq \int_{t_0'}^{t_1'} (L - \eta_c + \alpha(c))(d\xi(s), s)ds$$

*Department of Mathematics, Nanjing University, Nanjing 210093, China.
E-mail: chengcq@nju.edu.cn.

holds for each absolutely continuous curve ξ with $\xi(t_0') = \gamma(t_0)$, $\xi(t_1') = \gamma(t_1)$ and any $t_0 < t_1, t_0' < t_1'$ such that $t_0 - t_0', t_1 - t_1' \in \mathbb{Z}$. In this case, we call $d\gamma = (\gamma, \dot{\gamma})$ c-minimal (c-semi static) orbit. All c-minimal orbits constitute the Mañé set $\tilde{\mathcal{N}}(c)$, in which one can define the Mather set $\tilde{\mathcal{M}}(c)$ and the Aubry set $\tilde{\mathcal{A}}(c)$:

$$\tilde{\mathcal{M}}(c) \subseteq \tilde{\mathcal{A}}(c) \subseteq \tilde{\mathcal{N}}(c).$$

The symbols $\mathcal{M}(c)$, $\mathcal{A}(c)$ and $\mathcal{N}(c)$ denote their projection down to $M \times \mathbb{R}$ along each tangent fiber. Roughly speaking, Aubry set consists of the orbits of *stationary* motion, besides the orbits in the Aubry set, Mañé set contains *transient* orbits connecting different stationary states. For more details, one can refer the pioneer work of Mather [51, 52] as well as Mañé [57, 58]. There are also many contributions to this theory, for example see [4, 8, 10, 16, 26, 27, 34, 35, 36, 41, 59, 60].

The variational method has been developed a powerful tool for the study of global instability in Hamiltonian systems convex in action (slow) variables. Although the study of Arnold diffusion was started in the sixties of last century [1, 2], it was until the nineties before this problem was considered from variational point of view. In [12], a variational technique was applied to study the original example of Arnold. It seems the beginning.

By the study of recent years, great progress has been made towards solving the problem of Arnold diffusion by variational as well as geometric method. We mention the papers [29, 62, 9] and the announcements [64, 55]. It is almost impossible to list all works in this very active area, among which we also mention [15, 13, 33, 30, 31, 32, 37, 38, 39, 40, 44, 42, 43, 53], it is clearly incomplete.

In this lecture, we shall describe briefly, mainly based on [24, 25, 46], how to construct orbits connecting different Aubry sets by the variational method. The first version of [25] appeared in 2004. In general, along each orbit in Aubry set the variation of action (slow) variable is small. Along an orbit connecting different Aubry sets far from each other, action variable undergoes substantial variation. It implies diffusion.

To construct diffusion orbits, one needs to pose some hypotheses. Therefore, one of the important issues is to convince people the rationality of the hypotheses, for instance, the genericity. Up to now, to our knowledge, the generic property of the relevant hypotheses is proved only in [24, 25, 46], some cusp-residual property is provided in [62] for systems with two and half degrees of freedom.

2. Local Connecting Orbits

Throughout this report, we let $M = \mathbb{T}^n$. Let $\gamma\colon \mathbb{R} \to M$ be a solution of the Euler-Lagrange equation:

$$\frac{d}{dt}\Big(\frac{\partial L}{\partial \dot{x}}\Big) - \frac{\partial L}{\partial x} = 0.$$

We call the orbit $d\gamma = (\gamma, \dot{\gamma})$ connecting $\tilde{\mathcal{A}}(c)$ to $\tilde{\mathcal{A}}(c')$ if its α-limit set $\alpha(d\gamma) \cap \tilde{\mathcal{A}}(c) \neq \varnothing$ and its ω-limit set $\omega(d\gamma) \cap \tilde{\mathcal{A}}(c') \neq \varnothing$. A connecting orbit is called *global* if $\tilde{\mathcal{A}}(c')$ is not close to $\tilde{\mathcal{A}}(c)$, and called *local* if they are close to each other.

Global connecting orbits are constructed shadowing a sequence of local connecting orbits which are successively connected. Along each of these local connecting orbits the Lagrangian action attains local minimum. Up to now, we have found two types of local connecting orbits, i.e. type-c and type-h connecting orbits.

Type-h connecting orbit looks like heteroclinic orbit. Let $\pi : \bar{M} \to M$ be a finite covering manifold, $\tilde{\mathcal{N}}(c, \bar{M})$ denotes the corresponding Mañé set, then $\pi\tilde{\mathcal{N}}(c, \bar{M}) \supseteq \tilde{\mathcal{N}}(c)$, and in some case the inclusion is nontrivial. For example, if there is a neighborhood of some lower dimensional torus N containing the time-1-section of $\mathcal{N}(c)$: $N \supset \mathcal{N}_0(c)$, then $\pi\tilde{\mathcal{N}}(c, \bar{M}) \supsetneq \tilde{\mathcal{N}}(c)$ if $\bar{M}$ is chosen such that lift of N has two components $\bar{N}_1$ and $\bar{N}_2$. Indeed, it contains those homoclinic orbits along which the action attains minimum. For the topic of homoclinic orbits to Aubry sets, one can refer [14, 6, 28, 66]. In contrast with it, the Aubry set remains the same if we consider it in each finite covering manifold.

Theorem 2.1. *Let N be a neighborhood of lower dimensional torus such that the group $H_1(M, N, \mathbb{Z}) \neq 0$ and let $\bar{M}$ be a finite covering of M such that the lift of N has two components $\bar{N}_1$ and $\bar{N}_2$. Assume that, $N \supset \mathcal{N}_0(c)$ and $\pi\mathcal{N}_0(c, \bar{M})\backslash N$ contains an isolated point, then for each c' sufficiently close to c there exists an orbit connecting $\tilde{\mathcal{A}}(c)$ to $\tilde{\mathcal{A}}(c')$ or vise verse.*

One can refer [24, 25] for the proof. This theorem can only be applied to time-periodic systems. It has a version for autonomous systems [46]. The relevant conditions are required as $\pi\tilde{\mathcal{N}}(c, \bar{M})\backslash N \times \mathbb{R}^n$ contains an isolated orbit and $\alpha(c') = \alpha(c)$.

Type-c local connecting orbits are obtained by making use of the c-equivalence between corresponding cohomology classes. So-called c-equivalence is introduced in [52] first. However, the equivalence defined in [52] is of no interest for autonomous systems, as c is equivalent to c' if and only if c and c' are contained in the same flat of the α-function [7] and the Aubry set is the same for all cohomology classes in the interior of the flat [59]. Since the minimal points of the α-function constitute a flat, we only need to establish some kind of equivalence among those cohomology classes in a non-minimum level set of the α-function for autonomous systems.

Therefore, it is reasonable to assume that there exists a non-degenerate embedded $(n-1)$-dimensional torus $\Sigma_c \subset \mathbb{T}^n$ such that, restricted to the Mañé set $\mathcal{N}(c)$, the flow $\pi_x \phi_L^t$ is transversal to Σ_c (see [45]), where $\pi_x : TM \to M$ is the standard projection. We say Σ_c is a non-degenerate embedded $(n-1)$-dimensional torus if there is a smooth injection φ: $\mathbb{T}^{n-1} \to \mathbb{T}^n$ such that Σ_c is the image of φ, and the induced map φ_*: $H_1(\mathbb{T}^{n-1}, \mathbb{Z}) \to H_1(\mathbb{T}^n, \mathbb{Z})$ is an

injection. Let

$$V_c = \bigcap_U \{i_{U*} H_1(U, \mathbb{R}) : U \text{ is a neighborhood of} \mathcal{N}(c) \cap \Sigma_c\}.$$

Clearly,

$$V_c^{\perp} = \bigcup_U \{\ker i_U^* : U \text{ is a neighborhood of} \mathcal{N}(c) \cap \Sigma_c\}.$$

We say that $c, c' \in H^1(M, \mathbb{R})$ are *c-equivalent* if there exists a continuous curve Γ: $[0,1] \to H^1(M, \mathbb{R})$ such that $\Gamma(0) = c$, $\Gamma(1) = c'$, $\alpha(\Gamma(s))$ is constant for all $s \in [0,1]$, and for each $s_0 \in [0,1]$ there exists $\delta > 0$ such that $\Gamma(s) - \Gamma(s_0) \in V_{\Gamma(s_0)}^{\perp}$ whenever $s \in [0,1]$ and $|s - s_0| < \delta$.

From the definition of c-equivalence for autonomous systems, one can easily recover the original definition for time-periodic systems by considering t as an angle variable, $M \times \mathbb{T}$ as the configuration manifold and M as the codimension one section.

Theorem 2.2. ([24, 46]) *If c is equivalent to c', then there exists an orbit connecting $\tilde{\mathcal{A}}(c)$ to $\tilde{\mathcal{A}}(c')$.*

Applying this theorem to twist map, one immediately obtains the result in [50].

The proof of both theorem 2.1 and 2.2 are based on the semi-continuity of pseudo-connecting orbit set on parameters. We only discuss it for time-periodic case here ([25]), one can refer [46] for autonomous case. Let

$$L_{\eta,\mu,\psi} = L - \eta - \mu - \psi,$$

where η is a closed 1-form on M such that $[\eta] = c$, $\mu = \rho(t)\bar{\mu}$ in which $\bar{\mu}$ is a closed 1-form on M such that $[\bar{\mu}] = c' - c$, ρ is a smooth function of t with $\rho = 0$ for $t \le 0$ and $\rho = 1$ for $t \ge 1$; $\psi = \varrho(t)\bar{\psi}$ in which $\varrho = 0$ for $t \le 0$ as well as $t \ge 1$, $\bar{\psi}$ is a function depending on configuration coordinates only.

Definition 2.1. *The pseudo-connecting curve set $\mathscr{C}_{\eta,\mu,\psi}$ of $L_{\eta,\mu,\psi}$ consists of those absolutely continuous curves satisfying the condition*

$$\int_s^{\tau} L_{\eta,\mu,\psi}(d\gamma(t), t)dt = \inf_{\substack{s_1 - s \in \mathbb{Z},\ \tau_1 - \tau \in \mathbb{Z} \\ s_1 \le 0,\ \tau_1 \ge 1 \\ \xi(s_1) = \gamma(s) \\ \xi(\tau_1) = \gamma(\tau)}} \int_{s_1}^{\tau_1} L_{\eta,\mu,\psi}(d\xi(t), t)dt - (s_1 - s)\alpha(c) + (\tau_1 - \tau)\alpha(c').$$

holds for each $s \le 0$ and $\tau \ge 1$. The pseudo-connecting orbit set $\tilde{\mathscr{C}}_{\eta,\mu,\psi}$ is defined as

$$\tilde{\mathscr{C}}_{\eta,\mu,\psi} = \{d\gamma : \gamma \in \mathscr{C}_{\eta,\mu,\psi}\}.$$

If $N \subset M$ exists such that $\mathcal{A}(c) \subset N$, $\mathcal{A}(c') \subset N$ and $H_1(M, N, \mathbb{Z}) \neq 0$, the set $\mathscr{C}_{\eta,\mu,\psi,e_n}$ can be defined similarly to the set $\mathscr{C}_{\eta,\mu,\psi}$ by requiring an extra condition that $0 \neq [\gamma] \in H_1(M, N, \mathbb{Z})$ for each curve $\gamma \in \mathscr{C}_{\eta,\mu,\psi,e_n}$. Let

$$\tilde{\mathscr{C}}_{\eta,\mu,\psi,e_n} = \{d\gamma : \gamma \in \mathscr{C}_{\eta,\mu,\psi,e_n}\}.$$

Clearly, for each curve γ in the pseudo-connecting curve set, $\alpha(d\gamma) \subset \tilde{\mathcal{A}}(c)$ and $\omega(d\gamma) \subset \tilde{\mathcal{A}}(c')$ and $\cup_{t\in\mathbb{R},\gamma\in\mathscr{C}_{\eta,0,0}}(d\gamma(t), t) = \tilde{\mathcal{N}}(c)$. There would be an orbit connecting $\tilde{\mathcal{A}}(c)$ to $\tilde{\mathcal{A}}(c')$ if a curve in this set is a solution of the Euler-Lagrangian equation determined by L. The upper-semi continuity $(\eta, \mu, \psi) \to (\mathscr{C}_{\eta,\mu,\psi}, \mathscr{C}_{\eta,\mu,\psi,e_n})$ is used for the proof of both theorem 2.1 and 2.2. Refer [25] to see how to choose $\bar{\mu}$ and ψ when $\tilde{\mathcal{A}}(c)$ has isolated homoclinic orbits and when c is equivalent and close to c'.

The action along these local connecting orbits attains local minimum. Roughly speaking, we call a curve locally minimal if the action along this curve is smaller than the action along any other curve staying in its small neighborhood. This minimal property appears to be certain variational version of the transversal intersection of the "stable set" of an Aubry set with the "unstable set" of another Aubry set. Here is the precise definition for type-h. For $m_0, m_1 \in N$ and $H_1(M, N, \mathbb{Z}) \neq 0$

$$\begin{aligned} h^{t^-,t^+}_{\eta,\mu,\psi,e_n}(m_0, m_1) = &\inf_{\substack{\xi(-t^-)=m_0\\ \xi(t^+)=m_1\\ [\xi]\neq 0}} \int_{-t^-}^{t^+} L_{\eta,\mu,\psi}(d\gamma(s), s)ds \\ &+ t^-\alpha(c) + t^+\alpha(c'), \end{aligned}$$

$$h^{\infty}_{c}(m_0, m_1) = \liminf_{t\to\infty} \inf_{\substack{\xi(-t)=m_0\\ \xi(t)=m_1}} \int_{-t}^{t} (L - \eta)(d\xi(s), s)ds + 2t\alpha(c).$$

Local Minimal Property: *Assume $\mathcal{A}(c), \mathcal{A}(c') \subset N$, $H_1(M, N, \mathbb{Z}) \neq 0$, $[\eta] = c$ and $[\eta + \bar{\mu}] = c'$. There exist two open balls V^-, V^+ and two positive integers t^-, t^+ such that $\bar{V}^- \subset N\backslash\mathcal{M}_0(c)$, $\bar{V}^+ \subset N\backslash\mathcal{M}_0(c')$, $\gamma(-t^-) \in V_0$, $\gamma(t^+) \in V_1$ and*

$$\begin{aligned} &h^{\infty}_{c}(x^-, m_0) + h^{t^-,t^+}_{\eta,\mu,\psi,e_n}(m_0, m_1) + h^{\infty}_{c'}(m_1, x^+) \\ &- \liminf_{\substack{t_i^-\to\infty\\ t_i^+\to\infty}} \int_{-t_i^-}^{t_i^+} L_{\eta,\mu,\psi}(d\gamma(t), t)dt - t_i^-\alpha(c) - t_i^+\alpha(c') \\ &> 0 \end{aligned}$$

holds for any $(m_0, m_1) \in \partial(V_0 \times V_1)$, $x^- \in \mathcal{M}_0(c) \cap \pi_x(\alpha(d\gamma))$, $x^+ \in \mathcal{M}_0(c') \cap \pi_x(\omega(d\gamma))$. Where t_i^-, t_i^+ are the sequences such that $\gamma(-t_i^-) \to x^-$ and $\gamma(t_i^+) \to x^+$, and $\pi_x : TM \to M$ is the standard projection.

The local minimal property has a version for autonomous systems, refer [46] for the precise statement.

3. Global Connecting Orbits

The Aubry set for c can be connected to the Aubry set for c' by an orbit if there is a *generalized transition chain* joining c with c'. Such an orbit is a global connecting orbit shadowing a sequence of local connecting orbits, along which the Lagrangian action attains local minimum.

Definition 3.1. *Let $\pi : \bar{M} \to M$ be a finite covering of a compact manifold M and let c, c' be two cohomolgy classes in $H^1(M, \mathbb{R})$. We say that c is joined to c' by a generalized transition chain if there is a continuous curve Γ: $[0, 1] \to H^1(M, \mathbb{R})$ such that $\Gamma(0) = c$, $\Gamma(1) = c'$ and for each $\tau \in [0, 1]$ at least one of the following cases takes place:*

(I), there are finitely many Aubry classes, and there is a small $\delta_\tau > 0$ such that $\pi\mathcal{N}_0(\Gamma(\tau), \bar{M})\backslash(\mathcal{A}_0(\Gamma(\tau), M) + \delta_\tau) \neq \varnothing$ is totally disconnected;

(II), $\mathcal{N}_0(\Gamma(\tau), M)$ is homologically trivial, i.e. it has a neighborhood U_τ such that the inclusion map $H_1(U_\tau, \mathbb{R}) \to H_1(M, \mathbb{R})$ is the zero map.

This definition applies to time-periodic systems [25], one can find the version for autonomous systems from [46]. By the definition, there are finitely many cohomology classes $c_0 = c, c_1, \cdots, c_{k+1} = c'$ such that $\tilde{\mathcal{A}}(c_i)$ is connected to $\tilde{\mathcal{A}}(c_{i+1})$ either by type-h or by type-c connecting orbits.

To construct orbits connecting $\tilde{\mathcal{A}}(c)$ to $\tilde{\mathcal{A}}(c')$ in the time-periodic system, we introduce a modified Lagrangian

$$\tilde{L} = L - \eta_0 - \sum_{i=0}^{k} (-\tau_i)^* (\mu_i + \psi_i).$$

where τ_i represents a time translation operator such that $\tau^* u(\cdot, t) = u(\cdot, t + \tau)$, $[\eta_0 + \sum_{i=0}^{j} \bar{\mu}_i] = [\eta_{j+1}] = c_{j+1}$, μ_i and ψ_i are carefully chosen so that each curve either in $\mathscr{C}_{\eta_i, \mu_i, \psi_i, e_n}$ or in $\mathscr{C}_{\eta_i, \mu_i, \psi_i}$ is a solution of the Euler-Lagrange equation determined by L.

Given two points $m \in \mathcal{A}_0(c_0)$, $m' \in \mathcal{A}_0(c_{k+1})$, we consider the minimum of the action of $\tilde{L}$ along curves $\{\gamma : [-K, K' + \tau_k] \to M\}$

$$\inf \int_{-K}^{K'+\tau_k} \tilde{L}(d\gamma(t), t)dt + \sum_{i=1}^{k} (\tau_i - \tau_{i-1})\alpha(c_i) + K\alpha(c_0) + K'\alpha(c_{k+1})$$

under certain constraints. By choosing sufficiently large $\tau_{i+1} - \tau_i$, K and K', the minimizer does not touch the boundary of the constraints. Thus the minimizer is smooth everywhere, consequently, is the solution of the Euler-Lagrange equation for L. Let $K, K' \to \infty$, we obtain an orbit connecting $\tilde{\mathcal{A}}(c)$ to $\tilde{\mathcal{A}}(c')$. Refer [25] for details and refer [46] for the study in the autonomous case.

If the generic condition is dropped that there are finitely many Aubry classes for each cohomology class, there would be extra difficulty in the construction of global connecting orbits. In the definition of generalized transition chain, it is

required that the Aubry set has finitely many classes if it is connected to other Aubry set by an orbit of type-h. If the quotient Aubry sets occupy an interval (see the example in [56]), we do not know yet how to show the smoothness of the minimizer. It is proved in [54] the total disconnectedness of the quotient Aubry set in low dimensions.

It is not our final goal to produce somehow abstract framework such as a generalized transition chain, but it can be applied to interesting problems. The first non-trivial example is generic *a priori* unstable systems.

4. A Priori Unstable Systems

It has been shown in [24, 25, 46] that in *a priori* unstable systems a generalized transition chain exists generically in C^k-topology ($k = 3, 4, \cdots, \infty$). By *a priori* unstable system in time-periodic case we mean that it is the coupling of a rotator and a pendulum:

$$H(u, v, t) = h_1(p) + h_2(x, y) + P(u, v, t)$$

where $u = (q, x)$, $v = (p, y)$, $(p, q) \in \mathbb{R} \times \mathbb{T}$, $(x, y) \in \mathbb{T}^n \times \mathbb{R}^n$, P is a time-1-periodic small perturbation. $H \in C^r$ ($r = 3, 4, \cdots, \infty$) is assumed to satisfy the following hypothesis:

1, $h_1 + h_2$ is a convex function in v, i.e., the Hessian matrix $\partial^2_{vv}(h_1 + h_2)$ is positive definite. It is finite everywhere and has superlinear growth in v, i.e., $(h_1 + h_2)/\|v\| \to \infty$ as $\|v\| \to \infty$.

2, it is *a priori* hyperbolic in the sense that the Hamiltonian flow $\Phi^t_{h_2}$, determined by h_2, has a non-degenerate hyperbolic fixed point $(x, y) = (0, 0)$ and the function $h_2(x, 0) : \mathbb{T}^n \to \mathbb{R}$ attains its strict maximum at $x = 0$.

There is an invariant manifold $\Sigma \subset \mathbb{T}^{n+1} \times \mathbb{R}^{n+1} \times \mathbb{T}$ whose time-t-section Σ_t is a small deformation of a cylinder $\mathbb{T} \times \mathbb{R}$. Restricted to Σ_0, the time-1-map of the flow Φ^t_H is area-preserving and twist. By Aubry-Mather theory [48], the minimal invariant set $M_{\omega,0}$ for each rotation number ω is either an invariant curve, or an Denjoy set (cantori) or periodic points. Let $M_\omega = \cup_{t\in\mathbb{T}}(M_{\omega,0}, t)$.

To use the variational method we study it in Lagrangian formalism. Let

$$L(u, \dot{u}, t) = \max_v \langle v, \dot{u}\rangle - H(u, v, t),$$

it determines the coordinate transformation $\mathfrak{L}$: $(u, v, t) \to (u, \dot{u}, t)$. Obviously, L is a Tonelli Lagrangian, and in the space of $H^1(T^{1+n}, \mathbb{R}) = \mathbb{R}^{1+n}$ there is a channel $C \supset \mathbb{R} \times B_d$ such that for each $c \in C$ the Mather set $\tilde{\mathcal{M}}(c) = \mathfrak{L}M_\omega$ for certain $\omega \in \mathbb{R}$, where B_d is an n-dimensional ball with radius $d > 0$. Therefore, for each c of these cohomology classes, the homology group $H_1(M, \mathcal{N}_0(c), \mathbb{Z})$ is non-trivial.

Let Γ: $[0, 1] \to H^1(M, \mathbb{R})$ be a path connecting $c, c' \in C$ with $\Gamma(s) \in \mathbb{R} \times B_d$ for each $s \in [0, 1]$. Let $\sigma \subset [0, 1]$ be the set such that $s \in \sigma$ if an only if

$\mathcal{N}_0(\Gamma(s))$ is an invariant curve. Generically, σ is a Cantor set with positive Lebesgue measure. To obtain a generalized transition chain, we need to verify the condition that for each $s \in \sigma$ the set $\pi\mathcal{N}_0(\Gamma(s), \bar{M})\backslash\mathcal{N}_0(\Gamma(s)) + \delta \neq \varnothing$ is totally disconnected. The verification of this condition is by no means trivial, because the set σ is uncountable. Fortunately, it is done by checking some regularity of the barrier function with respect to s when they are restricted on σ. For other $s \in [0,1]\backslash\sigma$, the Manñé set $\mathcal{N}_0(\Gamma(s))$ is topologically trivial, i.e. it does not contain non-shrinkable circle. In this case, $\Gamma(s)$ is equivalent to any cohomology class nearby. Therefore, we can make use of the c-equivalence. In this way, we show that a generalized transition chain exists for C^k generic perturbation. Refer [46] for the relevant study on autonomous systems.

The *large gap* problem was considered as a major challenge for the construction of diffusion orbits. Corresponding to a strong resonant rotation number, generically there is a large Birkhoff instability region in the cylinder. By the methods already known, one is unable to verify whether the stable manifold of the invariant circle on one side intersects the unstable manifold of the circle on the other side, the gap between these two circles is too big (see [47]). It was proposed in [64] to overcome this problem by considering the Cantori in the regions so that the *stable* and *unstable sets* of the Cantori would bridge the large gap. The diffusion orbit claimed in [64] shadows a sequence of local connecting orbits of type-h only. In [24, 25, 46], the large gap problem is solved by constructing an orbit along a sequence of local connecting orbits with type-h as well as type-c. All these exhibit the power of the variational method.

5. Barrier Functions and Elementary Weak-KAM

To see that a Aubry set can be connected to another Aubry set nearby by a type-h orbit, we need to make sure that the minimal homoclinic orbits to the Aubry set are isolated, i.e. the minimal points of the barrier function in certain region or at certain section are isolated. As there are uncountably many barrier functions, this goal seems very difficult to reach. One strategy is to study the Hausdorff or box dimensions of the set of barrier functions.

Assume that the Aubry set is contained in N for all cohomology classes under consideration. Let $\mathscr{Z}$ be the set of those Lipschitz functions defined on $M\backslash N$, whose minimal point is not isolated, let $\mathscr{B}$ be the set of barrier functions restricted in $M\backslash N$. In vague language, $\mathscr{Z}$ is of "infinite co-dimensions". If the box dimension of $\mathscr{B}$ is finite, an open and dense set of small perturbations exists such that the set of barrier functions for each perturbed system is a translation of the original set $\mathscr{B} \to \mathscr{B} + u$ and $(\mathscr{B} + u) \cap \mathscr{Z} = \varnothing$, i.e. the minimal points are isolated for each barrier function.

The finiteness of box dimension is obtained if some Hölder regularity of barrier function on parameters is found. Up to now, we are only succeeded in

a priori unstable systems when the parameter is restricted on the Cantor set $\sigma \subset [0,1]$, corresponding to invariant circles. This is sufficient for the existence of a generalized transition chain obtained in [24, 25, 46]. If one try to obtain the chain composed by local connecting orbits of type-h only, he needs to consider the set of barrier functions for all $s \in [0,1]$ and to show the finiteness of its Hausdorff dimension. The verification appears unavailable yet by studying the regularity of the barrier functions on the parameter. However, the modulus of continuity for Peierls' barrier [49] is expected to be used for this purpose.

Barrier function can be expressed in terms of weak-KAM solutions [35]. However, we need simpler form of barrier function by so-called *elementary* weak-KAM solution [20]. It is possible when there are finitely many Aubry classes for each cohomology class, which is proved generic in [10].

We know that there is exactly only one weak-KAM if there is only one Aubry class. If there are finitely many Aubry classes for certain cohomology class c, denoted by $\{\mathcal{A}_{c,i},\ i = 1, 2, \cdots, k\}$ respectively, then we can construct small non-negative perturbation such that its support does not intersect with $\mathcal{A}_{c,i}$ which is the unique Aubry set for the perturbed system. Denoted by $u^-_{c,i,\epsilon}$ the unique weak-KAM, it is easy to see that there is exactly one weak-KAM $u^-_{c,i}$ such that $u^-_{c,i,\epsilon} \to u^-_{c,i}$. In the same way one can define $u^+_{c,i}$. We call them *elementary* weak-KAM. Thus, there are k pairs of elementary weak-KAM.

To construct type-h local connecting orbits, we need to consider the problem in certain finite covering configuration space. Let $u^{\pm}_{c,i}$ be the elementary weak-KAM with respect to the covering space, the elementary weak-KAM is nontrivial and the barrier function is defined as

$$B_{c,i,j} = u^-_{c,i} - u^+_{c,j}.$$

It measures the action along curves which joining the i-th Aubry class to j-th Aubry class in the covering space.

Generically, at most $n+1$ Aubry classes exist for each cohomology class, thus, there are finitely many barrier functions $B_{c,i,j}$ for each c. It implies that we obtained a set-valued map from the space of cohomology class to the space of Lipschitz functions. As the limit of a sequence of elementary weak KAM is also an elementary weak KAM, this set-valued map is upper semi-continuous when it is restricted to each closed set with the topology inherited from the standard topology in $\mathbb{R}^n$.

Recall that u^- is semi-concave and u^+ is semi-convex, we find that barrier function is differentiable at its local minimal points. By making use of the fact that each weak-KAM solution is a viscosity solution of the Hamilton-Jacobi equation, we are able to show in [20] that

The topological dimension of the set of barrier functions is finite.

Thus, it seems generically true for each cohomology that the stable set of the Aubry set does not coincide with unstable set of the Aubry set everywhere. Although it is useful in some special case, this is not sufficient to show the

existence of a generalized transition chain in generic systems. We need the following issue which remains open:

Open Problem 1: *Generically, the Hausdorff dimension of the set of barrier functions is finite.*

6. A Priori Stable Systems

A nearly integrable Hamiltonian system of KAM type is usually called *a priori* stable system when one studies the problem of Arnold diffusion:

$$H(x, y) = h(y) + h_\epsilon(x, y),$$

where h_ϵ is small in C^k-topology $\|h_\epsilon\|_k \leq \epsilon$ $(k \geq 3)$. We can write time-periodic system in the form of autonomous system by considering the time t as an extra angle variable. To study dynamical instability, one tries to construct global connecting orbits in resonant layers. Unlike KAM torus, a torus is easily destructed if the frequency is resonant.

A frequency ω is called k-resonant if there are exactly k independent integer vectors I_i such that $\langle \omega, I_i \rangle = 0$. In the unperturbed system, each standard torus $\{y = \text{constant}\}$ is invariant for the Hamiltonian flow. If the frequency on some torus is k-resonant, this torus has a foliation into a family of $(n-k)$-dimensional sub-tori, the frequency on these sub-tori is an $(n-k)$-dimensional and non-resonant. If the frequency on sub-torus is of Diophantine, some sub-tori survive small perturbations. Generically, if the frequency on n-torus is k-resonant, there are at least 2^k $(n-k)$-dimensional sub-tori surviving small perturbation [23], and some partial results were obtained earlier in [11, 22, 61]. Without generic assumption on perturbation, such result is proved in [17, 18] when $k = 1$, and it is shown in [19] one of the $(n-1)$-dimensional torus is the support of the relevant minimal measure. However, the existence of these lower dimensional tori depends on how small the perturbation is, comparing with the Diophantine coefficient. That is, for any perturbation in generic sense, there are always some Diophantine frequencies with "too small" coefficient such that the sub-tori are destructed by the perturbation. Also, a sub-torus is easily destructed if the frequency is of Liouville.

In practice, one is trying to construct global connecting orbits along a path $\Gamma\colon [0, 1] \to \mathbb{R}^n$ where each frequency $\partial_y h(\Gamma(s))$ is k-resonant with $k \geq n-2$. This path must pass through strong resonant points, around which the Hamiltonian is equivalent to the following [20]:

$$H'(x, y) = h'(y) + Z(x, y) + R(x, y)$$

where $\partial_y h'(\Gamma(s)) = \omega_s = (\omega_{1,s}, 0, \cdots, 0)$, Z is invariant to the flow $(x, y) \to (x + \omega_s t, y)$ and R is a higher order term comparing with Z in the following

sense: one can write R in the form

$$R = \sum_{|i|=0}^{k-2} Y_i(y - \Gamma(s))R_i(x, y),$$

where $Y_i(y)$ is i-homogenous in y and the following estimates hold:

$$\|Z\|_2 \leq 2\epsilon, \qquad \|R_i\|_2 \leq \epsilon^{1+(k-2-i)\lambda}, \quad \lambda = (3n+4)^{-1}.$$

To connect the Aubry set with the rotation vector ω_λ to other Aubry sets by the method we already know, we need to make sure that this Aubry set does not cover the whole configuration manifold, but this problem remains open when $n > 3$.

Open Problem 2: *Assume there is only one minimal measure μ_c for a given co-homology c, whose rotation vector is resonant, i.e. $\langle \rho(\mu_c), I\rangle = 0$ holds for some non-zero integer vector I. Is the relative homology group of M with respect to $\mathcal{A}(c)$ non-trivial, $H_1(M, \mathcal{A}(c), \mathbb{Z}) \neq 0$?*

In generic senese, this problem turns out to be simpler when $n = 3$. If we forget about the higher order term R, the Hamiltonian $h'(y) + Z(x, y)$ is independent of the first component of $x = (x_1, x_2, x_3)$ because it is invariant to the flow $(x, y) \to (x+\omega t, y)$. Thus, its dynamics is the same as a system with two degrees of freedom, $y_1 =$ constant. By the Lipschitz graph property of Aubry set, each orbit in Aubry set is either a periodic orbit or a fixed point provided the rotation vector satisfies some resonant condition. Under so-called cusp-residual assumptions, the Aubry set of the system $h' + Z + R$ remains in a small neighborhood of the periodic orbit or the fixed point. That is, $H_1(M, \mathcal{A}(c), \mathbb{Z}) \neq 0$ holds for every cohomology class under consideration.

However, even for three degrees of freedom the dynamical instability problem is not solved yet in cusp-residual *a priori* stable systems. Lack of estimation on the Hausdorff dimension of barrier function set, one has not yet found way to obtain the cusp-residual property. Nevertheless, we are optimistic about a weaker result:

Conjecture: *Given $\delta > 0$ and two points (x, y) (x', y') in an energy surface, there exists $\epsilon_0 > 0$ such that for each h_ϵ with $\|h_\epsilon\|_{C^k} \leq \epsilon < \epsilon_0$, there are uncountable many h'_ϵ with $\|h_\epsilon - h'_\epsilon\|_{C^2} < \epsilon^{1+(k-2)\lambda}$ such that the Hamiltonian flow of $H' = h + h'_\epsilon$ has a trajectory that visits the δ-neighborhood of (x, y) as well as that of (x', y').*

To see it, we note that H is approximated by $h + Z$ which is equivalent to a system with two degrees of freedom, $\|H - h - Z\|_{C^2} \leq \epsilon^{1+(k-2)\lambda}$. When the configuration manifold is a two-dimensional torus and the Mather set $\mathcal{M}(c)$ consists of one circle, the barrier function attains its minimum along each of its minimal homoclinic curve. Thus, the barrier function would be constant in an open set if the minimal homoclinic orbits are not isolated. Recall the

finiteness of the topological dimension of the barrier function set, it would be reasonable to expect that, generically, the minimal homoclinic orbits are isolated for each cohomology class. By choosing small and suitable perturbation R' to $h+Z \to h+Z+R'$, one can make sure that the minimal homoclinic orbits for $h + Z + R'$ are also isolated. That is, one obtains a generalized transition chain.

The conjecture of *topological instability* in higher-dimensional case by Arnold [2, 3] is expected to be true for exact Hamiltonian systems, that is, the flow is determined by a Hamiltonian function. It is interesting to recall a result for volume-preserving diffeomorphism: a set of co-dimension one tori with positive Lebesque measure survives perturbations (see [21, 65, 63]). Consequently, topological instability is not a generic phenomenon for volume-preserving diffeomorphisms.

Acknowledgement

The author is partially supported by National Basic Research Program of China (973 Program, 2007CB814800), National Natural Science Foundation of China (Grant 10531050) and Basic Research Program of Jiangsu Province, China (BK2008013).

References

[1] Arnol'd V. I., *Instability of dynamical systems with several degrees of freedom,* (Russian, English) Sov. Math., Dokl., **5** (1964), 581–585; translation from Dokl. Akad. Nauk. SSSR, **156** (1964), 9–12.

[2] Arnol'd, V.I., *Small denominators and problems of stability of motion in classical and celestial mechanics,* Russ. Math. Survey **18** (1963), 85–192.

[3] Arnol'd V.I., *Dynamical Systems,* III. Encyclopaedia of Mathematical Sciences, **3** Springer-Verlag Berlin Heidelberg, (1988)

[4] Bangert V., *Minimal geodesics*, Ergod. Theory Dynam. Syst. **10** (1990), 263–286.

[5] Bangert V., *Geodesic rays, Busemann functions and monotone twist maps*, Cal. Vari. PDE **2** (1994) 49–63.

[6] Bernard P., *Homoclinic orbits to invariant sets of quasi-integrable exact maps*, Ergod. Theory Dynam. Syst. **20** (2000), 1583–1601.

[7] Bernard P., *Connecting orbits of time dependent Lagrangian systems,* Ann. Inst. Fourier (Grenoble), **52(5)**(2002), 1533–1568.

[8] Bernard P., *Symplectic aspects of Mather theory*, Duke Math. J. **136** (2007), 401–420.

[9] Bernard P., *The dynamics of pseudographs in convex Hamiltonian systems*, J. Amer. Math. Soc. **21** (2008), 615–669.

[10] Bernard P. and Contreras G., *A generic property of families of lagrangian systems*, Annals of Math. **167** (2008), 1099–1108.

[11] Berstein D. and Katok A., *Birkhoff periodic orbits for small perturbations of completely inetgrable Hamiltonians*, Invent. Math. **88** (1984), 225–142.

[12] Bessi U., *An approach to Arnold's diffusion through the calculus of variations,* Nonlinear Anal., **26**(1996), 1115–1135.

[13] Bessi U., Chierchia L. and Valdinoci E., *Upper bounds on Arnold diffusion times via Mather theory,* J. Math. Pures Appl., **80(1)**(2001), 105–129.

[14] Bolotin S. *Homoclinic orbits in invariant tori of Hamiltonian systems,* Dynamical systems in classical mechanics, 21–90, Amer. Math. Soc. Transl. Ser. 2, **168**, Amer. Math. Soc., Providence, RI, 1995.

[15] Bourgain J. and Kaloshin V., *On diffusion in high-dimensional Hamiltonian systems*, J. Functional Analysis, **229** (2005), 1–61.

[16] Carneiro M.J.D.,*On minimizing measures of the action of autonomous Lagrangians*, Nonlinearity **6** (1996), 1077–1085.

[17] Cheng, C.-Q., *Birkhoff-Kolmogorov-Arnold-Moser tori in convex Hamiltonian systems*, Commun. Math. Phys. **177** (1996), 529–559.

[18] Cheng, C.-Q., *Lower dimensional invariant tori in the regions of instability for nearly integrable Hamiltonian systems*, Commun. Math. Phys. **203** (1999), 385–419.

[19] Cheng C.-Q., *Minimal invariant tori in the resonant regions of nearly integrable Hamiltonian systems*, Trans. Amer. Math. Soc. **357** (2005), 5067–5095.

[20] Cheng C.-Q., *Abundance of Arnold diffusion in a priori stable systems*, in preparation.

[21] Cheng C.-Q. and Sun Y.S., *Existence of invariant tori in three dimensional measure-preserving mappings*, Celestial Mechanics and Dynamical Astronomy **45** (1989/1990), 275–292.

[22] Cheng C.-Q. and Sun Y.S., *Existence of periodically invariant curves in three dimensional measure-preserving mappings*, Celestial Mechanics and Dynamical Astronomy **45** (1989/1990), 293–303.

[23] Cheng C.-Q. and Wang S.L., *The surviving of lower dimensional tori from resonance torus of Hamiltonian systems*, J. Diff. Eqns. **155** (1999), 311–326.

[24] Cheng C.-Q. and Yan J., *Existence of diffusion orbits in a priori unstable Hamiltonian systems,* J. Diff. Geometry **67** (2004), 457–517.

[25] Cheng C.-Q. and Yan J., *Arnold diffusion in Hamiltonian Systems: a priori unstable case,* mp_arc (2004) 04-265, J. Diff. Geometry **82** (2009), 229–277.

[26] Contreras G., Delgado J. and Iturriaga R., *Lagrangian flows: the dynamics of globally minimizing orbits II*, Bol. Soc. Bras. Mat. **28** (1997), 155–196.

[27] Contreras G. and Paternain G. P., *Connecting orbits between static classes for generic Lagrangian systems,* Topology, **41** (2002), 645–666.

[28] Cui X. Cheng C.-Q. and Cheng W., *Existence of infinitely many homoclinic orbits to Aubry sets for positive definite Lagrangian systems*, J. Diff. Eqns. **214** (2005), 176–188.

[29] Delshams A., de la Llave R. and Seara T. M., *Geometric mechanism for diffusion in Hamiltonian systems overcoming the large gap problem: heuristic and rigorous verification of a model*, Memoirs Amer. Math. Soc. **179**(844), (2006).

[30] Delshams A. de la Llave R. and Seara T. M., *Orbits of unbounded energy in quasi-periodic perturbations of geodesic flows*, Advances in Mathematics **202** (2006), 64–188.

[31] Delshams A. de la Llave R. and Seara T. M., *Geometric properties of the scattering map of a normally hyperbolic invariant manifold*, Advances in Mathematics **217** (2008), 1096–1153.

[32] Delshames A. and Huguet G., *Geography of resonances and Arnold diffusion in a priori unstable Hamiltonian systems*, Nonlineairty **22** (2009), 1997–2077.

[33] Douady R., *Stabilité ou instabilité des points fixes elliptiques*, Ann. Sci. École Norm. Sup. (4) **21** (1988), 1–46.

[34] Fathi A., *Théorème KAM faible et théorie de Mather sue les systèmes*, C. R. Acad. Sci. Paris Sér. I Math. **324** (1997), 1043–1046.

[35] Fathi A., *Weak KAM Theorem in Lagrangian Dynamics*, Cambridge Studies in Adavnced Mathematics, Cambridge University Press, (2009).

[36] Fathi A. and Mather J., *Failure of convergence of the Lax-Oleinik semi-group in the time-periodic case,* Bull. Soc. Math. France, **128** (2000), 473–483.

[37] Fontich E. & Martin P., *Arnold diffusion in perturbations of analytic integrable Hamiltonian systems*, Discret Contin. Dynam. Systems, **7** (2001), 61–84.

[38] Gidea M. and de la Llave R., *Topological methods in the instability problem of Hamiltonian systems*, Discrete and Continuous Dynamical Systems **14** (2006), 294–328.

[39] Gidea M. and Robinson C., *Shadowing orbits for transitive chains of invariant tori alternating with Birkhoff zone of instability*, Nonlinearity **20** (2007), 1115–1143.

[40] Gidea M. and Robinson C., *Obstruction argument for transition chains of tori interspersed with gaps*, Discrete and Continuous Dynamical Systems series S, **2** (2009), 393–416.

[41] Iturriaga R., *Minimizing measures for time dependent Lagrangians*, Proc. London Math. Soc. **73** (1996), 216–240.

[42] Kaloshin V. and Levi M., *An example of Arnold diffusion for near-integrable Hamiltonians*, Bulletin Amer. Math. Soc. **45** (2008) 409–427.

[43] Kaloshin V. and Levi M., *Geometry of Arnold diffusion*, SIAM Review **50** (2008) 702–720.

[44] Kaloshin V. Mather J. and Valdinoci E., *Instability of resonant totally elliptic points of symplectic maps in dimension 4*, Asterisque **297** (2004) 79–116.

[45] Li X. *On c-equivalence*, Science in China, series A, **52** (2009), 2389–2396.

[46] Li X. and Cheng C-Q., *Connecting orbits of autonomous Lagrangian systems*, Nonlinearity **23** (2010), 119–141.

[47] Lochak P., *Arnold diffusion: a compendium of remarks and questions*, in Hamiltonian systems with three or more degrees of freedom, Kluwer Acad. Publ., Dordrecht (1999), 168–183.

[48] Mather J., *Existence of quasiperiodic orbits for twist homeomorphisms of the annulus,* Topology, **21**(1982), 457–467.

[49] Mather J., *Modulus of continuity for Peierls's barrier*, in Periodic Solutions of Hamiltonian Systemsand Related Topics, *edited by P.H.Rabinowitz et al*, D. Reidel Publishing Company (1987), 177–202.

[50] Mather J., *Variational construction of orbits of twist diffeomorphisms*, J. Amer. Math. Soc. **4** (1991), 207–263.

[51] Mather J., *Action minimizing invariant measures for positive definite Lagrangian systems,* Math. Z., **207** (1991), 169–207.

[52] Mather J., *Variational construction of connecting orbits,* Ann. Inst. Fourier (Grenoble), **43**(1993), 1349–1386.

[53] Mather J., *Variational construction of Trajectories for time periodic Lagrangian systems on the two torus*, (1997), (unfinished manuscript).

[54] Mather J., *Total disconnectedness of the quotient Aubry set in low dimensions*, Commun. Pure. Appl. Math. **56** (2003), 1178–1183.

[55] Mather J., *Arnold diffusion, I: Announcement of results*, J. Math. Sciences **124** (2004), 5275–5289.

[56] Mather J., *Exmaples of Aubty sets*, Ergod. Th. Dynam. Sys. **24** (2004), 1667–1723.

[57] Mañé R., *Lagrangian flows: the dynamics of globally minimizing orbits*, Proceedings Int. Congress in Dynamical Systems (*Montevideo 1995*), Pitman Research Notes in Math. **362** (1996), 120–131.

[58] Mañé R., *Generic properties and problems of minimizing measures of Lagrangian systems*, Nonlinearity, **9** (1996), 273–310.

[59] Massart D., *On Aubry sets and Mather's action functional*, Israel J. Math. **134** (2003) 157–171.

[60] Osuna O., *Vertices of Mather's beta function*, Ergod. Theory Dynam. Syst. **25** (2005) 949–955.

[61] Treschev D.V., *The mechanism of destruction of resonance tori of Hamiltonian systems*, Math USSR Sbornik **68** (1991), 181–203.

[62] Treschev D.V., *Evolution of slow variables in a priori unstable Hamiltonian systems*, Nonlinearity, **17** (2004), 1803–1841.

[63] Xia Z.H., *Existence of invariant tori in volumn-preserving diffeomorphisms*, Ergod. Th. & Dynam. Syst. **12** (1992), 621–631.

[64] Xia Z.H., *Arnold diffusion: a variational construction,* Proceedings of the International Congress of Mathematicians, Doc. Math. 1998, Extra Vol. II, 867–877.

[65] Yoccoz J.C., *Travaux de Herman sur les tores Invariants*, Asterisque **206** (1992), 311–344.

[66] Zheng Y. and Cheng C.-Q., *Homoclinic orbits of positive definite Lagrangian systems*, J. Diff. Eqns. **229** (2006) 297–316.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Generic Dynamics of Geodesic Flows

Gonzalo Contreras*

Abstract

We present some perturbation methods which help to describe the generic dynamical behaviour of geodesic flows.

Mathematics Subject Classification (2000). Primary 53D25; Secondary 37D40.

Keywords. Geodesic flows, topological entropy, twist map, closed geodesic.

The difficulties in studying generic properties of geodesic flows with respect to other classes of dynamical systems are twofold. The obvious difficulty of making perturbations for the geodesic equations and the fact that perturbations of geodesic flows are never local.

Indeed, let (M, g) be a compact riemannian manifold and write $g = \sum_{ij} g_{ij}(x)\, dx_i \otimes dx_j$. The phase space of the geodesic flow is the unit tangent bundle SM. A perturbation of the coefficients of the riemannian metric $g_{ij}(x)$ with support $A \subset M$ changes the geodesic vector field along the whole interior of the fiber $SA = \pi^{-1}(A)$, where $\pi : SM \to M$ is the projection. All the known proofs of the closing lemma (cf. [25]) use local perturbations and hence they can not be applied to geodesic flows.

Similar difficulties arise when one tries to change the Euler-Lagrange flow of a lagrangian $L : TM \to \mathbb{R}$ on a given energy level with perturbations by a potential i.e. $L'(x, v) = L(x, v) + \psi(x)$, where $\psi : M \to \mathbb{R}$ is a function on M. For a mechanical lagrangian, this corresponds to perturbing the conditions of the problem without changing Newton's law.

In [7] we prove Theorem 1 below, here we describe the ingredients its proof.

Theorem 1. *On any closed manifold M with $\dim M \geq 2$ the set of C^∞ riemannian metrics whose geodesic flow contains a non-trivial hyperbolic basic set is open and dense in the C^2 topology.*

That basic set is a horseshoe obtained from a homoclinic point. Using symbolic dynamics, the existence of a horseshoe implies that such geodesic flows

*Partially supported by CONACYT Mexico.
CIMAT, P.O. Box 402, 36000 Guanajuato GTO, Mexico. E-mail: gonzalo@cimat.mx.

have positive topological entropy and that the number of closed geodesics grow exponentially with their length.

1. Bumpy Metrics

The simplest invariant set in a geodesic flow is a periodic orbit $\Gamma = (\gamma, \dot{\gamma})$ arising from a closed geodesic γ on M. Given a small transversal section Σ to Γ at $\Gamma(0)$ in SM define the *Poincaré map* $\mathcal{P} = \mathcal{P}(\Sigma, \gamma) : \Sigma \hookleftarrow$ as the first return map under the geodesic flow. Its derivative $P = d_{\Gamma(0)}\mathcal{P}$ is called the *linearized Poincaré map*. The geodesic γ is said *non-degenerate* if 1 is not an eigenvalue of P. This is the necessary condition in the implicit function theorem to obtain a continuation of Γ under perturbations of the metric. It is also equivalent to Γ being a non-degenerate critical point of the action functional $A = \int \|v\|^2$ with appropriate normalizations[1]. A metric is said *bumpy* if all its periodic orbits are non-degenerate.

Let $\mathcal{R}^r(M)$, be the set of C^r riemannian metrics on M endowed with the C^r topology. The *Bumpy Metric Theorem* states that the set of bumpy metrics contains a residual subset in $\mathcal{R}^r(M)$, $2 \leq r \leq \infty$. It was first announced by Abraham [1] but the first complete proof was given by Anosov [2]. Klingenberg and Takens [19] made a useful improvement:

Write $n+1 = \dim M$. Given an integer $k \geq 1$ let $J^k_s(n)$ be the set of k-jets of smooth symplectic maps of $(\mathbb{R}^{2n}, 0) \hookleftarrow$. A set $Q \subset J^k_s(n)$ is said *invariant* if for all $\sigma \in J^k_s(n)$, $\sigma\, Q\, \sigma^{-1} = Q$. The theorem in [19] proves that if Q is a residual and invariant subset of $J^k(n)$ then the set of metrics such that the Poincaré map of every closed geodesic is in Q contains a residual set in $\mathcal{R}^r(M)$. See [30], [29] for analogous theorems on hypersurfaces of $\mathbb{R}^{n+2}$.

The proof in [19] is based on a local perturbation theorem which says that if γ is a closed geodesic for $g \in \mathcal{R}^r(M)$ there is $g' \in \mathcal{R}^r(M)$ arbitrarily close to g such that γ is a closed geodesic for g' and its Poincaré map belongs to Q. This implies the theorem above provided that the set of closed geodesics is countable. This condition is ensured by the case $k = 1$ proved by Anosov [2] together with the Bumpy Metric Theorem.

2. Twist Maps

We say that a closed geodesic is *hyperbolic* if its linearized Poincaré map has no eigenvalues of modulus 1 (in a transversal section inside SM). We say that it is *elliptic* if it is non-degenerate and non-hyperbolic.

The existence of a generic elliptic periodic orbit gives dynamical information about the geodesic flow. If it is partially elliptic, i.e. if not all eigenvalues have

[1] i.e. on the space of closed curves in M with fixed parametrization interval $[0, \ell]$ and initial point in the transversal section Σ.

modulus 1, using [15] one can obtain an invariant central manifold N where the Poincaré map $\mathcal{P}|_N$ is totally elliptic and N is normally hyperbolic.

Imposing generic conditions specifying only the jets of the Poincaré maps at the periodic points [7, §3] it is possible apply Klingenberg and Takens Theorem [19] to obtain coordinates in which a restriction of the Poincaré map $\mathcal{P}|_N$ becomes a weakly monotonous exact twist map on $\mathbb{T}^q \times \mathbb{R}^q$ which is C^1 near a totally integrable twist map. In this conditions we have the Birkhoff-Lewis theorem (see Moser [18, appendix 3.3]) which says that any punctured neighbourhood of the elliptic point contains a periodic point.

Indeed, the condition to write the Birkhoff normal form are that the elliptic points are *4-elementary*, this is that the eigenvalues of modulus one $\rho_1, \ldots, \rho_q$; $\overline{\rho_1}, \ldots, \overline{\rho_q}$ satisfy

$$\prod_{i=1}^{q} \rho_i^{\nu_i} \neq 1 \quad \text{whenever} \quad 1 \leq \sum_{i=1}^{q} |\nu_i| \leq 4. \tag{1}$$

Then the normal form is $\mathcal{P}(x, y) = (X, Y)$, where

$$\begin{aligned} Z_k &= e^{2\pi i\, \phi_k}\ z_k + g_k(z), \\ \phi_k(z) &= a_k + \textstyle\sum_{\ell=1}^{q} \beta_{k\ell}\, |z_\ell|^2 \end{aligned}$$

$z = x+iy$, $Z = X+iY$, $\rho_i = e^{2\pi\, i\, a_k}$ and $g(z) = g(x, y)$ has vanishing derivatives up to order 3 at the origin. We say that the normal form is *weakly monotonous* if the matrix $\beta_{k\ell}$ is non-singular. The property $\det \beta_{k\ell} \neq 0$ is independent of the particular choice of normal form. In these coordinates, the matrix $\beta_{k\ell}$ can be detected from the 3-jet of $\mathcal{P}$ at $\theta = (0, 0)$ and it can be seen that the property $\{\, (1)$ and $\det \beta_{k\ell} \neq 0 \,\}$ is open and dense in the jet space $J^3_s(q)$. Changing the coordinates to $(\theta, r) \in \mathbb{T}^q \times \mathbb{R}^q$, where $z_j = \sqrt{\varepsilon\, r_j}\, e^{2\pi i \theta_j}$ on $r_j > 0, \quad \forall j$, the Poincaré map becomes a weakly monotonous exact twist map of $\mathbb{T}^q \times \mathbb{R}^q$. We restrict our discussion to the generic set of riemannian metrics all of whose closed geodesics are 4-elementary and have weakly monotonous normal forms.

Moreover, using techniques developed by Arnaud [3] we prove in [7] that $\mathcal{P}|_N$ has a 1-elliptic periodic point. This is a periodic point whose linearized Poincare map on a transversal Σ inside SM has exactly two eigenvalues of modulus 1. Such a periodic point has a normally hyperbolic central manifold where the Poincaré map is an exact twist map of the 2-dimensional annulus $\mathbb{S}^1 \times \mathbb{R}$.

Such a generic twist map contains periodic points for all rational rotation numbers in an interval. In fact for any such rational rotation number there are elliptic and hyperbolic periodic points which have homoclinic intersections [20].

3. The Kupka-Smale Theorem

A single homoclinic intersection in a geodesic flow can be made transversal by a perturbation argument by Donnay [11] in dimension 2 and Petroll [23] in higher

dimensions. But perhaps this is not enough to make transversal two invariant manifolds.

Another argument that can be used to change invariant manifolds or single orbits in geodesic flows and also in lagrangian systems with perturbations by a potential can be made along the following lines. Weak stable and weak unstable manifolds are lagrangian submanifolds for the canonical symplectic form. A lagrangian submanifold contained in a level set of an autonomous hamiltonian is invariant under the hamiltonian flow. Then it is enough to deform the stable manifold W^s to a lagrangian submanifold Λ which is transversal to W^u and then perturb the metric so that the geodesic hamiltonian $H|_\Lambda$ is constant. We will have that $W^s = \Lambda$ for the new geodesic flow. The details appear in [9, Theorem 2.5 and Appendix A].

This argument together with Anosov-Klingenberg-Takens theorem gives

Theorem 2. *Let $Q \subset J_s^{k-1}(n)$ be residual and invariant. There is a residual subset $\mathcal{G} \subset \mathcal{R}^k(M)$ such that if $g \in \mathcal{G}$ then*

- *The $(k-1)$-jet of the Poincaré map of every closed geodesic of g is in Q.*
- *All heteroclinic intersections of hyperbolic orbits of g are transversal.*

Choosing Q in the previous theorem as the condition $\{\,(1) \text{ and } \det \beta_{k\ell} \neq 0\,\}$ as above, we have that for a Kupka-Smale geodesic flow, if it contains an elliptic closed geodesic then it has a transversal homoclinic orbit and hence a hyperbolic subset. It remains to study the case in which all closed geodesics are hyperbolic.

4. Many Closed Geodesics

Bangert [5], Hingston [14] and Rademacher [26], [28] prove that a C^k generic riemannian metric, $2 \leq k \leq \infty$, contains infinitely many closed geodesics.

If the geodesic flow contains a generic elliptic closed geodesic, this is implied by the Birkhoff-Lewis theorem (Moser [18, appendix 3.3]). But in this case Rademacher [28] obtains infinitely many closed geodesics by imposing only conditions on the 1-jet of the Poincaré map, which is easier to perturb as in Anosov [2]. If there are finitely many closed geodesics Rademacher obtains a resonance condition on the average indices of the geodesics. If there is one elliptic closed geodesic, its average index can be perturbed to break the resonance and hence obtain infinitely many closed geodesics. If all closed geodesics are hyperbolic, then Hingston [14, §6.1] and Rademacher [26, Theorem 1] prove that there are infinitely many.

It is not known if a simply connected manifold can have all its closed geodesics hyperbolic. In [27] Rademacher proves that in the examples of ergodic geodesic flows in S^2 of Donnay [10] and Burns-Gerber [6], all the homologically visible closed geodesics are hyperbolic.

5. Stable Hyperbolicity

In order to prove the generic existence of a homoclinic orbit and hence a hyperbolic set when all closed geodesics are hyperbolic, we use the theory of stable hyperbolicity developed by Mañé [22].

Recall that a linear map $T : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is *hyperbolic* if it has no eigenvalue of modulus 1. Equivalently, if there is a splitting $\mathbb{R}^{2n} = E^s \oplus E^u$ and $M \in \mathbb{Z}^+$ such that $T(E^s) = E^s$, $T(E^u) = E^u$, $\|T^M|_{E^s}\| < \frac{1}{2}$, $\|T^{-M}|_{E^u}\| < \frac{1}{2}$.

Let $Sp(n)$ be the group of symplectic linear isomorphisms of $\mathbb{R}^{2n}$. We say that a sequence $\xi : \mathbb{Z} \to Sp(n)$ is *periodic* if there is $m \geq 1$ such that $\xi_{m+i} = \xi_i$ for all $i \in \mathbb{Z}$. A periodic sequence is said *hyperbolic* if the linear map $\prod_{i=1}^m \xi_i$ is hyperbolic. In this case the stable and unstable subspaces of $\prod_{i=1}^m \xi_{j+i}$ are denoted by $E^s_j(\xi)$ and $E^u_j(\xi)$ respectively.

A family $\xi = \{\xi^\alpha\}_{\alpha\in\mathcal{A}}$ of sequences in $Sp(n)$ is *bounded* if there exists $Q > 0$ such that $\|\xi^\alpha_i\| < Q$ for every $\alpha \in \mathcal{A}$ and $i \in \mathbb{Z}$. Given two families of periodic sequences in $Sp(n)$, $\xi = \{\xi^\alpha\}_{\alpha\in\mathcal{A}}$ and $\eta = \{\eta^\alpha\}_{\alpha\in\mathcal{A}}$, we say that they are *periodically equivalent* if they have the same indexing set $\mathcal{A}$ and for all $\alpha \in \mathcal{A}$ the periods of ξ^α and η^α coincide. Given two periodically equivalent sequences in $Sp(n)$, $\xi = \{\xi^\alpha\}_{\alpha\in\mathcal{A}}$ and $\eta = \{\eta^\alpha\}_{\alpha\in\mathcal{A}}$ define

$$d(\xi, \eta) = \sup\{\, \|\xi^\alpha_i - \eta^\alpha_i\| \,:\, \alpha \in \mathcal{A},\ i \in \mathbb{Z}\,\}.$$

We say that a family ξ is *hyperbolic* if for all $\alpha \in \mathcal{A}$ the periodic sequence ξ^α is hyperbolic. We say that a hyperbolic periodic family ξ is *stably hyperbolic* if there is $\varepsilon > 0$ such that any periodically equivalent family η satisfying $d(\xi, \eta) < \varepsilon$ is also hyperbolic.

Finally, we say that a family of periodic sequences ξ is *uniformly hyperbolic* if there exist $K > 0$, $0 < \lambda < 1$ and subspaces $E^s_i(\xi^\alpha)$, $E^u_i(\xi^\alpha)$, $\alpha \in \mathcal{A}$, $i \in \mathbb{Z}$ such that $\xi_j(E^\tau_j(\xi^\alpha)) = E^\tau_{j+1}(\xi^\alpha)$ for all $\alpha \in \mathcal{A}$, $j \in \mathbb{Z}$, $\tau \in \{s, u\}$ and

$$\left\| \prod_{i=1}^m \xi^\alpha_{j+i} \Big|_{E^s_j(\xi^\alpha)} \right\| < K\,\lambda^m \quad \text{and} \quad \left\| \left(\prod_{i=1}^m \xi^\alpha_{j+i} \Big|_{E^u_j(\xi^\alpha)} \right)^{-1} \right\| < K\,\lambda^m$$

for all $m \in \mathbb{Z}^+$, $\alpha \in \mathcal{A}$, $j \in \mathbb{Z}$.

In [7] we prove

Theorem 3. *If ξ^α is a bounded stably hyperbolic family of periodic sequences of symplectic linear maps then it is uniformly hyperbolic.*

6. The Perturbation Lemma

Let Γ be a set of closed geodesics. Construct a family ξ of periodic sequences in $Sp(n)$ given by the linearized time 1 maps of the geodesic flow restricted to the normal bundle $\mathcal{N}$ in SM to the geodesic vector field in $\Gamma \subset SM$.

Suppose that there are infinitely many closed geodesics in Γ and that the family ξ is uniformly hyperbolic. Then the subspaces E^s, E^u are continuous in Γ and hence they can be extended continuously to the closure $\overline{\Gamma}$. The closure would be a uniformly hyperbolic set. By the Spectral Decomposition Theorem it contains a non-trivial hyperbolic basic set because it is not a union of isolated periodic orbits.

We say that a set Γ of closed geodesics for a metric g_0 is *stably hyperbolic* if there is a neighbourhood $\mathcal{U} \subset \mathcal{R}^2(M)$ of g_0 in the C^2-topology such that for every $g \in \mathcal{U}$, the analytic continuation $\Gamma(g)$ of Γ exists and all the orbits in $\Gamma(g)$ are hyperbolic. In [9] we prove a perturbation lemma which implies that if Γ is a stably hyperbolic set of closed geodesics then the corresponding family ξ of symplectic linear maps is stably hyperbolic.

Using that result we obtain that a geodesic flow can either be perturbed to contain a generic elliptic closed geodesic, and hence a twist maps and homoclinics, or the set of its periodic orbits is stably hyperbolic and then contains a hyperbolic basic set.

We need a lemma in which one can perturb the linearized Poincaré map of the time one map in single geodesic in Γ by a fixed amount independently of the length, position, self-intersection or self-accumulation of the geodesic. Since there is no "transversal space" to mitigate the perturbation, such a lemma can only hold in the C^2 topology. The lemma is written for a small geodesic segment.

For simplicity we assume that all our riemannian metrics have injectivity radius larger than 2. Due to an algebraic obstruction in our proof of the lemma we have to assume that the initial metric g_0 is in the set $\mathcal{G}_1$ of metrics with the property that every geodesic segment of length $\frac{1}{2}$ has one point where the sectional curvatures are all different. In [7, §6 and appendix A] we prove that $\mathcal{G}_1$ is C^2 open and C^∞ dense.

For the sequel we need to characterize $\mathcal{G}_1$ precisely. The orthogonal group $\mathcal{O}(n)$ acts on the set of symmetric matrices $\mathcal{S}(n) \subset \mathbb{R}^{n\times n}$ by conjugation: $K \mapsto Q\,K\,Q^*$, $K \in \mathcal{S}(n)$, $Q \in \mathcal{O}(n)$. Given $g \in \mathcal{R}^2(M)$, define the map $K_g : SM \to \mathcal{S}(n)/\mathcal{O}(n)$ as $K_g(\theta) := [K]$, where $K_{ij} = \langle R_g(\theta, e_i)\,\theta,\, e_j\rangle$, R_g is the curvature tensor of g and $\{\,\theta, e_1, \ldots, e_n\,\}$ is a g-orthonormal basis for $T_{\pi(\theta)}M$. Let $h : \mathcal{S}(n)/\mathcal{O}(n) \to \mathbb{R}$ be the function

$$h([K]) = \prod_{1\le i<j\le n} (\lambda_i - \lambda_j)^2,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of K. Let $H : \mathcal{R}^2(M) \to [0, +\infty[$ be

$$H(g) = \min_{\theta\in SM} \max_{t\in[0,\frac{1}{2}]} h(K_g(\theta)).$$

Then $\mathcal{G}_1 = \{\, g \in \mathcal{R}^2(M) \,:\, H(g) > 0\,\}$.

Fix a C^∞ riemannian metric g_0 on M. Let $\gamma : [0,1] \to M$ be a geodesic segment for g_0. Let W be any neighbourhood of $\gamma([0,1])$ in M. Let $\mathcal{F} =$

$\{\eta_1, \ldots, \eta_m\}$ be any finite set of geodesic segments defined on $[0,1]$ with the following properties

- The endpoints of η_i are not contained in W.
- The segment $\gamma([0,1])$ intersects each η_i transversally.

Let U be a neighbourhood of $\cup\mathcal{F} := \cup_{i=1}^{m}\eta_i([0,1])$. Denote by $\mathcal{R}^\infty(g_0, \gamma, \mathcal{F}, W, U)$ the set of C^∞ riemannian metrics g on M for which γ is a geodesic segment, $g = g_0$ on $\gamma([0,1])$ and $g = g_0$ on $U \cup (M \setminus W)$.

Let $\mathcal{N}_t = \{\zeta \in T_{\dot\gamma(t)}SM \,|\, \langle d\pi(\zeta), \dot\gamma(t)\rangle = 1\}$ be the subspace transversal to the geodesic vector field given by the kernel of the Liouville 1-form $\lambda_{(x,v)}(\zeta) = \langle \zeta, v\rangle_x$. This subspace is the same for all metrics g with $g = g_0$ on $\gamma([0,1])$. Fix symplectic orthonormal basis for $\mathcal{N}_0$ and $\mathcal{N}_1$. Identify these subspaces with $\mathbb{R}^{2n}$ and the symplectic linear maps $\mathcal{N}_0 \to \mathcal{N}_1$ with $Sp(n)$. Let ϕ_t^g be the geodesic flow of a metric g and $S : \mathcal{R}^\infty(g_0, \gamma, \mathcal{F}, W, U) \to Sp(n)$ be the linearized Poincaré map $S(g) := d_{\dot\gamma(0)}\phi_1^g|_{\mathcal{N}_0}$. Write

$$B(g,\gamma) := \max\left\{ \|g|_\gamma\|_{C^4}\,,\ \left[\max_{t\in[\frac14,\frac34]} h(K_g(\dot\gamma(t)))\right]^{-1} \right\}.$$

Theorem 4. *Let $g_0 \in \mathcal{G}_1 \cap \mathcal{R}^r(M)$, $4 \le r \le \infty$. Given a neighbourhood $\mathcal{U} \subset \mathcal{R}^2(M)$ of g_0 there is $\delta = \delta(B(g_0,\gamma),\mathcal{U}) > 0$ such that given any γ, W and $\mathcal{F}$ as above there is a neighbourhood $U = U(B(g_0,\gamma),\mathcal{U},\gamma,W,\mathcal{F})$ of $\cup\mathcal{F}$ in M such that the image of $\mathcal{G}_1 \cap \mathcal{U} \cap \mathcal{R}^\infty(g_0,\gamma,\mathcal{F},W,U)$ under the map S contains the ball of radius δ centered at $S(g_0)$.*

The actual perturbation is made in a small neighbourhood of one point in $\gamma([\frac14,\frac34])$, so that the theorem can be applied independently to adjacent segments. In order to perturb the linearized Poincaré map on a periodic orbit we use Theorem 4 sequentially on segments of the orbit, taking care that the support of the perturbations are disjoint, as suggested in figure 1. The radius δ can remain constant $\delta = \delta(\max\{\|g_0\|_{C^4}, H(g_0)^{-1}\},\mathcal{U})$ in this process. Because despite the C^4 norm of the perturbed metric grows and $H(g)$ changes, the estimate of δ only depends on the bounds along the segment γ and subsequent perturbations have disjoint supports. Each perturbation is C^2 small and all perturbations remain in $\mathcal{U}$.

The statement of the lemma for a closed orbit is as follows. Given a closed geodesic γ for g_0 and a neighbourhood W of γ let $\mathcal{R}^r(g_0,\gamma,W)$ be the set of C^r riemannian metrics g for which γ is a geodesic, and such that $g = g_0$ on $\gamma \cup (M \setminus W)$. Let T be the minimal period of γ and let $m \in \mathbb{N}$ and $\tau \in [\frac12, 1]$ be such that $m\tau = T$. Let $\gamma_k(t) = \gamma(t + k\tau)$, $t \in [0,\tau]$ and for $g \in \mathcal{R}^r(g_0,\gamma,W)$ let $S_k(g) = d_{\gamma_k(0)}\phi_\tau^g \in Sp(n)$, identifying symplectic linear maps $\mathcal{N}(\dot\gamma_k(0)) \to \mathcal{N}(\dot\gamma_k(\tau))$ with $Sp(n)$, $\mathcal{N}(\theta) = \ker\lambda_\theta|_{T_\theta SM}$.

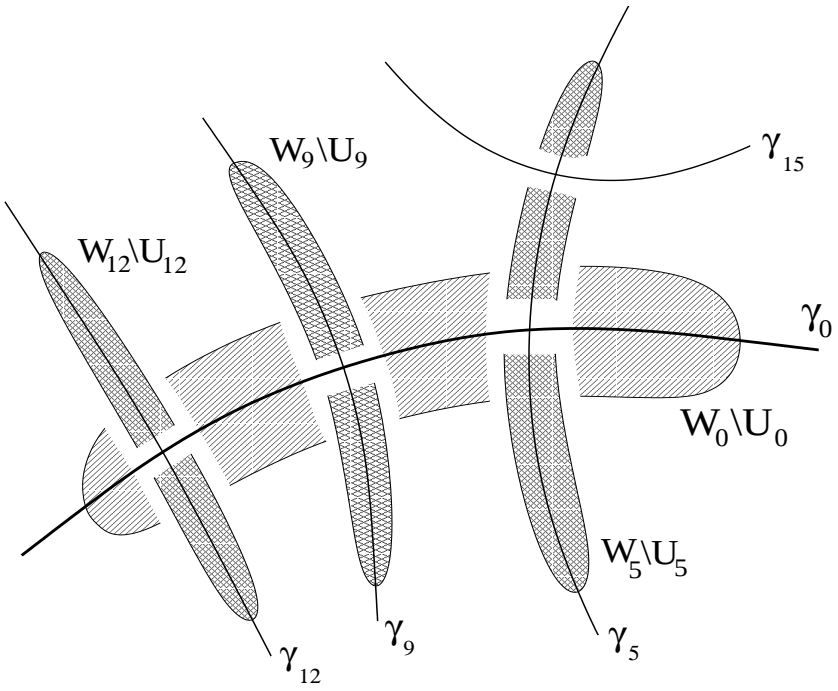


Figure 1. Avoiding self-intersections.

Corollary. *Let $g_0 \in \mathcal{G}_1 \cap \mathcal{R}^r(M)$, $4 \leq r \leq \infty$. Given a neighbourhood $\mathcal{U}$ of g_0 in $\mathcal{R}^2(M)$, there exists $\delta = \delta(g_0, \mathcal{U}) > 0$ such that if $g \in \mathcal{U}$, γ is a cosed geodesic for g_0 and W is a tubular neighbourhood of γ, then the image of $\mathcal{U} \cap \mathcal{G}_1 \cap \mathcal{R}^r(\gamma, g_0, W) \to \Pi_{k=0}^{m-1} Sp(n)$, under the map $(S_0, \ldots, S_{m-1})$, contains the product of balls of radius δ centered at $S_k(g_0)$ for $0 \leq k < m$.*

The derivative of the geodesic flow is represented by Jacobi fields. To prove Theorem 4 one has to perturb the solutions of the Jacobi equation. The Jacobi equation is difficult to solve but the perturbation of the Jacobi equation giving the derivative $d_g S$ can be solved by variation of parameters in terms of the original solution $S(g)$. This allows to estimate the expansion of $d_g S$ and then the radius δ.

7. Elliptic Geodesics in the Sphere

Another problem in which these methods have been used [8] is to prove that there is a C^2 open and dense set of riemannian metrics in $\mathbb{S}^2$ which contain an elliptic closed geodesic.

Henri Poincaré [24] claimed that every convex surface in $\mathbb{R}^3$ contains an elliptic closed geodesic. But Grjuntal [13] showed a counterexample. Pinching conditions to obtain an elliptic closed geodesic on spheres have been given in Grjuntal [12], Thorbergsson [31] and Ballmann, Thorbergsson, Ziller [4].

If the metric can not be perturbed to a metric with an elliptic closed geodesic, then its set of closed orbits is stably hyperbolic and then its closure is uniformly hyperbolic. The geodesic flow of $\mathbb{S}^2$ can not be Anosov, because Anosov geodesic flows do not have conjugate points [17], [21].

The geodesic flow is the Reeb flow of the Liouville contact form on SM. The unit tangent bundle of $\mathbb{S}^2$ is $\mathbb{RP}^3$. Its double cover is $\mathbb{S}^3$ and the geodesic flow of $\mathbb{S}^2$ lifts to the Reeb flow of a tight contact form on $\mathbb{S}^3$. If the metric is bumpy one can apply the theory of Hofer, Wysocki, Zehnder [16].

In the dynamically convex case, there is a surface of section which is a disk transversal to all but one orbit of the Reeb flow which is the boundary of the disk. The return map to the disk preserves the finite area form which is the differential of the contact form. This leads to a contradiction because it can be proved that a homoclinic class of an area preserving map which is not Anosov can not be uniformly hyperbolic. In the non-dynamically convex case we use geometric arguments on the finite energy foliation of [16] to get a contradiction.

References

[1] Ralph Abraham, *Bumpy metrics*, Global Analysis, Proc. Sympos. Pure Math. vol. XIV (S.S. Chern and S. Smale, eds.), 1970, pp. 1–3.

[2] Dmitri Victorovich Anosov, *On generic properties of closed geodesics.*, Izv. Akad. Nauk SSSR, Ser. Mat. **46** (1982), 675–709, Eng. Transl.: Math. USSR, Izv. **21**, 1–29 (1983).

[3] Marie-Claude Arnaud, *Type des points fixes des difféomorphismes symplectiques de* $\mathbf{T}^n \times \mathbf{R}^n$, Mém. Soc. Math. France (N.S.) (1992), no. 48, 63.

[4] Werner Ballmann, Gudlaugur Thorbergsson, and Wolfgang Ziller, *Closed geodesics on positively curved manifolds*, Ann. of Math. (2) **116** (1982), no. 2, 213–247.

[5] Victor Bangert and Nancy Hingston, *Closed geodesics on manifolds with infinite abelian fundamental group*, J. Differ. Geom. 19 (1984) 277–282.

[6] Keith Burns and Marlies Gerber, *Real analytic Bernoulli geodesic flows on* S^2, Ergodic Theory Dynam. Systems **9** (1989), no. 1, 27–45.

[7] Gonzalo Contreras, *Geodesic flows with positive topological entropy, twit maps and hyperbolicity*, to appear in Ann. Math.

[8] Gonzalo Contreras and Fernando Oliveira, C^2*-densely, the 2-sphere has an elliptic closed geodesic*, Ergodic Theory Dynam. Systems **24** (2004), no. 5, 1395–1423, Michel Herman's memorial issue.

[9] Gonzalo Contreras and Gabriel Paternain, *Genericity of geodesic flows with positive topological entropy on* S^2, Jour. Diff. Geom. **61** (2002), 1–49.

[10] Victor J. Donnay, *Geodesic flow on the two-sphere. I. Positive measure entropy*, Ergodic Theory Dynam. Systems **8** (1988), no. 4, 531–553.

[11] ———, *Transverse homoclinic connections for geodesic flows.*, Hamiltonian dynamical systems: history, theory, and applications. Proceedings of the international conference held at the University of Cincinnati, OH (USA), March 1992. IMA Vol. Math. Appl. **63**, 115–125 (H. S. Dumas et al., ed.), New York, NY: Springer-Verlag, 1995.

[12] A. I. Grjuntal′, *The existence of a closed nonselfintersecting geodesic of general elliptic type on surfaces that are close to a sphere*, Mat. Zametki **24** (1978), no. 2, 267–278, 303.

[13] ———, *The existence of convex spherical metrics all of whose closed nonselfintersecting geodesics are hyperbolic*, Izv. Akad. Nauk SSSR Ser. Mat. **43** (1979), no. 1, 3–18, 237.

[14] Nancy Hingston, *Equivariant Morse theory and closed geodesics.*, J. Differ. Geom. **19** (1984), 85–116.

[15] Morris W. Hirsch, Charles C. Pugh, and Michael Shub, *Invariant manifolds*, Springer-Verlag, Berlin, 1977, Lecture Notes in Mathematics, Vol. 583.

[16] Helmut Hofer, Krzysztof Wysocki, and Eduard Zehnder, *Finite energy foliations of tight three-spheres and Hamiltonian dynamics*, Ann. of Math. (2) **157** (2003), no. 1, 125–255.

[17] Wilhelm Klingenberg, *Riemannian manifolds with geodesic flow of Anosov type*, Ann. of Math. (2) **99** (1974), 1–13.

[18] ———, *Lectures on closed geodesics*, Grundlehren der Mathematischen Wissenschaften, Vol. 230, Springer-Verlag, Berlin-New York, 1978.

[19] Wilhelm Klingenberg and Floris Takens, *Generic properties of geodesic flows*, Math. Ann. **197** (1972), 323–334.

[20] Patrice Le Calvez, *Étude Topologique des Applications Déviant la Verticale*, Ensaios Matemáticos, **2**, Sociedade Brasileira de Matemática, Rio de Janeiro, 1990.

[21] R. Mañé, *On a theorem of Klingenberg.*, Dynamical systems and bifurcation theory, Proc. Meet., Rio de Janeiro/Braz. 1985, Pitman Res. Notes Math. Ser. 160, 319–345 (1987)., 1987.

[22] Ricardo Mañé, *An ergodic closing lemma*, Ann. of Math. (2) **116** (1982), no. 3, 503–540.

[23] Dietmar Petroll, *Existenz und Transversalitaet von homoklinen und heteroklinen Orbits beim geodaetischen Fluss* , Univ. Freiburg, Math. Fak. 42 S., 1996.

[24] Henri Poincaré, *Sur les lignes géodésiques des surfaces convexes*, Trans. Amer. Math. Soc. **6** (1905), no. 3, 237–274.

[25] Charles C. Pugh and Clark Robinson, *The C^1 closing lemma, including Hamiltonians*, Ergodic Theory Dyn. Syst. **3** (1983), 261–313.

[26] Hans-Bert Rademacher, *On the average indices of closed geodesics.*, J. Differ. Geom. **29** (1989), no. 1, 65–83.

[27] ———, *Morse theory and closed geodesics. (morse-theorie und geschlossene geodätische.)*, Ph.D. thesis, Bonner Mathematische Schriften. 229. Bonn: Univ. Bonn, 111 p., 1992.

[28] ———, *On a generic property of geodesic flows.*, Math. Ann. **298** (1994), no. 1, 101–116.

[29] Luchezar Stojanov and Floris Takens, *Generic properties of closed geodesics on smooth hypersurfaces.*, Math. Ann. **296** (1993), no. 3, 385–402 (English).

[30] Luchezar N. Stojanov, *A bumpy metric theorem and the Poisson relation for generic strictly convex domains.*, Math. Ann. **287** (1990), no. 4, 675–696 (English).

[31] Gudlaugur Thorbergsson, *Non-hyperbolic closed geodesics.*, Math. Scand. **44** (1979), 135–148.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Applications of Measure Rigidity of Diagonal Actions

Manfred Einsiedler*

Abstract

Furstenberg and Margulis conjectured classifications of invariant measures for higher rank actions on homogeneous spaces. We survey the applications of the partial measure classifications result by Einsiedler, Katok, and Lindenstrauss to number theoretic problems.

Mathematics Subject Classification (2010). Primary 37A45; Secondary 37D40, 11J13, 11J04.

Keywords. Invariant measures, entropy, homogeneous spaces, Littlewood's conjecture, diophantine approximation on fractals, distribution of periodic orbits, ideal classes, divisibility in integer Hamiltonian quaternions.

1. Introduction

The interaction between the theory of dynamical systems and number theory, and in particular of the theory of diophantine approximation, has a long and fruitful history. In particular, the study of the action of subgroups $H < \mathrm{SL}_n(\mathbb{R})$ on the quotient $X_n = \mathrm{SL}_n(\mathbb{Z})\backslash \mathrm{SL}_n(\mathbb{R})$ is often intimately linked to number theoretic problems.

For instance G. A. Margulis used in the late 1980's the subgroup

$$\mathrm{SO}(2,1)(\mathbb{R})^\circ \subset \mathrm{SL}_3(\mathbb{R})$$

acting on

$$X_3 = \mathrm{SL}_3(\mathbb{Z})\backslash \mathrm{SL}_3(\mathbb{R})$$

by right translation to prove the long-standing Oppenheim conjecture concerning the values $Q(\mathbb{Z}^n)$ of an indefinite quadratic form in $n \geq 3$ variables, see [38].

*This research has been supported by the NSF (0554373) and the SNF (200021-127145).
ETH Zürich, Departement Mathematik, Rämistrasse 101, 8092 Zürich, Switzerland. E-mail: manfred.einsiedler@math.ethz.ch.

Here the acting group $\mathrm{SO}(2,1)(\mathbb{R})^\circ$ is a simple non-compact subgroup of $\mathrm{SL}_3(\mathbb{R})$ that is generated by unipotent one-parameter subgroups. Here a *unipotent one-parameter subgroup* is the image of a homomorphism $u : \mathbb{R} \to \mathrm{SL}_n(\mathbb{R})$ given by $u(t) = \exp(tm)$ for $t \in \mathbb{R}$ and some given nilpotent matrix $m \in \mathrm{Mat}_n(\mathbb{R})$.

Due to the work of Ratner [44, 45] the dynamics of H on X_n is to a large extent understood if H is generated by unipotent one-parameter subgroups. These theorems and their extensions by Dani, Margulis, Mozes, Shah, and others, have found numerous applications in number theory and dynamics. We refer to [32], [40], and [46] for more details on these important topics.

These notes concern the dynamics of the diagonal subgroup A of $\mathrm{SL}_n(\mathbb{R})$, with the aim to explain the many connections between number theory and the action of A on X_n (or similar actions). We hope that the compilation of these applications will serve as a motivation to find new connections.

Before we list the applications let us briefly describe the dynamics of the diagonal subgroup. First we need to point out that the dynamical properties of a one-parameter subgroup $a(t)$ of A is quite different from the dynamical properties of a unipotent one-parameter subgroup. For instance if $n = 2$ then the dynamical system given by right translation of the diagonal elements

$$a(t) = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} = \mathrm{diag}(e^{t/2}, e^{-t/2})$$

on X_2 is precisely the geodesic flow on the unit tangent bundle of the modular surface $M = \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. This flow is hyperbolic and one can find, e.g. by using the Anosov shadowing lemma, an abundance of arbitrarily weird orbits. We refer to [31] for the theory of hyperbolic flows and to [18, §9.7] for a discussion of A-invariant measures on X_2. This should be contrasted with the dynamics of the horocycle flow, i.e. the dynamics of the unipotent one-parameter subgroup $u(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ on X_2 where every orbit is either periodic or is equidistributed in X_2 with respect to the Haar (or Liouville) measure, see [4] and [18, §11.7].

However, if $n \geq 3$ and A denotes the full $(n-1)$-dimensional subgroup, then it is expected that the orbits are better behaved. For instance, we have the following conjecture of G. A. Margulis.

Conjecture 1.1. *Let $n \geq 3$ and let*

$$A = \{\mathrm{diag}(a_1, \ldots, a_n) : a_1, \ldots, a_n > 0, a_1 \cdots a_n = 1\}.$$

Then any $x \in X_n = \mathrm{SL}_n(\mathbb{Z})\backslash \mathrm{SL}_n(\mathbb{R})$ for which xA has compact closure in X_n must actually belong to a periodic (i.e. compact) orbit.

The problem of classifying all A-invariant measures on X_n for $n \geq 3$ is strongly related to the study of orbits[1]. In fact, by the pointwise ergodic theorem the time-average of a function over the orbit of a (typical) point approximates the integral of the function with respect to an invariant measure. Here is the analogous conjecture for invariant measures. Here is the analogous conjecture for invariant measures, due to Margulis, and Katok and Spatzier.

Conjecture 1.2. *Let $n \geq 3$ and let A be as above. Then any A-invariant and ergodic probability measure on X_n is necessarily the normalized Haar measure on a finite volume orbit xH of an intermediate group $A \subseteq H \subseteq \mathrm{SL}_n(\mathbb{R})$.*

However, we also would like to mention the simplest case of such a conjectured classification result. Furstenberg proved in [21] that the full torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ and certain finite sets of rational points are the only closed sets in $\mathbb{T}$ that are invariant under $x \to 2x$ and $x \to 3x$. The related question for invariant measures is a famous conjecture also due to Furstenberg (unpublished).

Conjecture 1.3. *Let μ be an invariant and ergodic probability measure on $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ for the joint action of $x \to 2x$ and $x \to 3x$. Then either μ equals the Lebesgue measure or must have finite support (consisting of rational numbers).*

These conjectures and their counterparts on similar homogeneous spaces are still open, we refer to [14] for related more general versions of this conjecture. What is known towards Conjecture 1.2 is the following theorem which we obtained in joint work with A. Katok and E. Lindenstrauss [12].

Theorem 1.4. *Let $n \geq 3$. Then an A-invariant and ergodic probability measure μ on X_n either equals the normalized Haar measure on a closed finite volume orbit xH of an intermediate group $A \subset H \subseteq \mathrm{SL}_n(\mathbb{R})$ or the measure-theoretic entropy $h_\mu(a)$ vanishes for all $a \in A$. If n is a prime number, then necessarily $H = \mathrm{SL}_n(\mathbb{R})$.*

This theorem is the analogue to the theorem of Rudolph [48] towards Conjecture 1.3. Moreover, it is related to works of A. Katok, Spatzier, and Kalinin [29, 30, 27, 28] and uses arguments both from our joint work with A. Katok [11] and the paper of E. Lindenstrauss [36] on the Arithmetic Quantum Unique Ergodicty conjecture. We do not want to describe the history of the theorem in detail and instead refer to [14].

Theorem 1.4 reduces the problem to understanding the case where entropy is zero. Depending on the application, this unsolved problem is avoided by showing that the measure in the application has positive entropy. However, this sometimes (but not always) forces extra conditions in the application. In these cases the theorem in the application would improve if one could show

[1] Conjecture 1.2 implies Conjecture 1.1, but in contrast to the case of unipotent dynamics an equidistribution result for A-orbits is not conjectured.

that an ergodic measure with vanishing entropy must be the volume measure on a periodic A-orbit xA.

Theorem 1.4 has been generalized (technically speaking to all maximal $\mathbb{R}$ resp. $\mathbb{Q}_p$-split diagonal subgroups acting on any quotient of an S-algebraic group). However, for this care must be taken as e.g. no such theorem can be true for the two-parameter diagonal subgroup A on the space $\mathrm{SL}_2(\mathbb{Z}) \times \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R}) \times \mathrm{SL}_2(\mathbb{R})$, as any product of two invariant measures for the geodesic flow would be an invariant measure for A. Moreover, this scenario can hide e.g. inside $\Gamma\backslash \mathrm{SL}_4(\mathbb{R})$. Whether or not this is an issue crucially depends on the lattice Γ, which stands in contrast to the theorems concerning subgroups generated by unipotent subgroups where the precise nature of the lattice is not that important. We refer to [14] for the precise formulation and more details.

Here is the list of applications that we will discuss.

- Arithmetic Quantum Unique Ergodicity, see §2
- Diophanine approximation for points (and vectors) in fractals, see §3
- Non-uniformity of bad approximations of $n\alpha$, see §4
- Littlewood's conjecture, see §5
- Compact orbits and ideal classes, see §6
- Counting rational points in a certain variety, see §7
- Divisility properties of Hamiltonian quaternions, see §8

We also want to refer to the lecture notes [15] for the Clay summer school in Pisa in 2007 which explain in detail the (otherwise not so readily available) background of the papers [11, 12] as well as their content, and discusses two applications. Finally, we also wrote together with E. Lindenstrauss a joint survey [14] on this topic, which explains in detail the general conjectures and partial measure classifications and again some of the applications. In contrast to these surveys and lecture notes, we want to give here a description of all the applications and try to point out most concretely how these topics are connected to diagonal actions. For these applications we do not have to consider the most general theorems as all applications concern quotients of the group SL_n (or products of the form $\mathrm{SL}_n \times \cdots \times \mathrm{SL}_n$). This is unfortunate, as the theorems (in appropriate formulations) are more general.

I would like to thank my co-authors A. Katok, E. Lindenstrauss, Ph. Michel, and A. Venkatesh for the many collaborations on these subjects.

2. Arithmetic Quantum Unique Ergodicity

Historically the first application of a partial measure classification for diagonal flows (outside of ergodic theory) concerns the distributional properties of Hecke-Maass cusp forms ϕ on $M = \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ and similar quotients of the hyperbolic plane $\mathbb{H}$ by congruence subgroups. Here a Maass cusp form is a smooth function

ϕ on M which is an eigenfunction of the hyperbolic Laplace operator Δ_M and also belongs to $L^2(M)$ — we will always assume the normalization $\|\phi\|_2 = 1$. A Hecke-Maass cusp form is a Maass cusp form that in addition is also an eigenfunction of the Hecke operators T_p for all p.

Rudnick and Sarnak [47] conjectured that for any sequence of Maass cusp forms ϕ_i on M for which the eigenvalues go to infinity the probability measures defined by $|\phi_i|^2 \operatorname{dvol}_M$ converges in the weak* topology to the uniform measure dvol_M. These conjectures are of interest to mathematical physics as well as number theory. In quantum physics eigenfunctions of Δ are energy states of a free (spinless, non-relativistic) quantum particle, moving in the absence of external forces on M. In number theory the eigenfunctions are of central importance due to many connections between them and the theory of L-functions. We refer to the survey [50] and the more recent [49] for more details.

After a conditional proof of the following theorem by Watson [59] relying on the generalized Riemann hypothesis, Lindenstrauss [36] used a partial measure classification to show the equidistribution except for the possibility that the limit measure may not be a probability measure (or even the zero measure). Soundararajan [56] complemented the proof of Lindenstrauss showing that the limit measure must indeed be a probability measure. Together this gives the following unconditional theorem.

Theorem 2.1 (Arithmetic Quantum Unique Ergodicity). *Let $M = \Gamma\backslash\mathbb{H}$, with Γ a congruence lattice over $\mathbb{Q}$. Then $|\phi_i|^2 \operatorname{dvol}_M$ converges in the weak* topology to dvol_M as $i \to \infty$ for any sequence of Hecke-Maass cusp forms for which the Maass eigenvalues $\lambda_i \to -\infty$ as $i \to \infty$.*

The connection of this problem to the problem of classifying invariant measures on $X = \Gamma\backslash \operatorname{SL}_2(\mathbb{R})$ with respect to the geodesic flow is well motivated due to the interpretation of the Maass forms on M as the distribution of quantum particles on the surface M with a given energy (which up to a constant equals the eigenvalue for the Laplace operator) and the study of the semi-classical limit (corresponding to the limit where the energy goes to infinity). Moreover, Shnirelman [55], Zelditch [61] and Coin-de Verdeire [6] used this connection before to show the so-called Quantum Ergodicity. This theorem says that for any compact quotient $\Gamma\backslash\mathbb{H}$ and a subsequence of all eigenfunctions of density one, the measures $|\phi_i|^2 \operatorname{dvol}_M$ indeed converge to dvol_M. Part of this proof is the construction of a so-called micro-local lift of a weak* limit, which is a measure μ on the unit tangent bundle $X = \Gamma\backslash \operatorname{SL}_2(\mathbb{R})$ that is invariant under the geodesic flow.

The additional assumption in Theorem 2.1 that ϕ_i is also an eigenfunction of the Hecke-operators can be used to prove additional properties of the micro-local lift. Indeed, Bourgain and Lindenstrauss [2] show that a micro-local lift must have the property that all of its ergodic components have positive entropy. Here the positivity of entropy is shown by proving that the measure of a small ball $B_\epsilon(x)$ for $x \in X$ decays like $\ll_x \epsilon^{1+\delta}$ for $\delta = \frac{2}{9}$. The 'trivial bound' in

this case is $\ll \epsilon$ since the measure is known to be invariant under the one-dimensional subgroup A consisting of diagonal elements. Any improvement of the form $\ll_x \epsilon^{1+\delta}$ for some $\delta > 0$ shows positivity of entropy of almost all ergodic components.

Furthermore, Lindenstrauss [36] also shows that such a micro-local lift of a sequence of Hecke-Maass cusp forms has an additional recurrence property under the p-adic group $\mathrm{SL}_2(\mathbb{Q}_p)$ for any p — this is a much weaker requirement than invariance but suffices due to the following theorem [36].

Theorem 2.2. *Let Γ be a congruence lattice over $\mathbb{Q}$, let $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ and let μ be a probability measure satisfying the following properties:*

(I) μ is invariant under the geodesic flow,

(E) the entropy of every ergodic component of μ is positive for the geodesic flow, and

(R) μ is Hecke p-recurrent for a prime p.

Then μ is the uniform Haar measure m_X on X.

The proof of this theorem uses an idea from [11] and also an idea from the work of Ratner on the rigidity of the horocycle flow [42, 43]. The latter is surprising as the measure μ under consideration has a-priori very little structure with respect to the horocycle flow.

We refer to the lecture notes [17], which explain carefully the arguments in [2] and [36] (with the exception of the proof of Theorem 2.2).

After the work of Lindenstrauss, Silberman and Venkatesh [53] have generalized this approach to quotients of $\mathrm{SL}_n(\mathbb{R})$ by congruence lattices arising from division algebras, where the degree n of the division algebra is assumed to be a prime number. (In this case Theorem 1.4 holds in the same way.)

3. Diophantine Approximation for Points in Fractals

The connection between the continued fraction expansion and the geodesic flow on $X_2 = \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ goes back to work of Artin [1], see also [51, 52]. This link between Diophantine approximation of real numbers and dynamics on homogeneous spaces has been extended to higher dimension by Dani [3] and since then has been used successfully by many authors. We will recall this connection below.

In the theory of metric Diophantine approximations, one wishes to understand how well vectors in $\mathbb{R}^d$ can be approximated by rational vectors. In particular, we say $\mathbf{v} \in \mathbb{R}^d$ is *well approximable* if for any $c > 0$ there are infinitely

many nonzero integers q for which there exists an integer vector $\mathbf{p} \in \mathbb{Z}^d$ with

$$\left\| \mathbf{v} - \frac{1}{q}\mathbf{p} \right\| \leq \frac{c}{q^{1+\frac{1}{d}}}.$$

Similarly we say $\mathbf{v}$ is *badly approximable* if there exists a constant $c > 0$ such that

$$\left\| \mathbf{v} - \frac{1}{q}\mathbf{p} \right\| \geq \frac{c}{q^{1+\frac{1}{d}}}. \tag{1}$$

for all $q \in \mathbb{Z}$ and $\mathbf{p} \in \mathbb{Z}^d$. We will write WA (resp. BA) for the set of well approximable (resp. badly approximable) vectors. It is well known that almost every $\mathbf{v}$ is well approximable, but that the set of badly approximable vectors is also in many ways big — e.g. W. Schmidt has shown that the set of badly approximable vectors has full Hausdorff dimension.

Recently, the question how special submanifolds or fractals within $\mathbb{R}^d$ intersect the set of badly or well approximable vectors (as well as other classes of vectors with special Diophantine properties) has attracted attention. For instance, it was shown for the Cantor set $C \subset [0,1]$ in [33] and [34], that the dimension of $C \cap \mathrm{BA}$ is full, i.e. equals $\log 2/\log 3$. However, until recently little was known about the intersection of WA with fractals. In joint work [10] with L. Fishman and U. Shapira we obtained the following application of Theorem 2.2.

Theorem 3.1. *Almost any point in the middle third Cantor set (with respect to the natural measure) is well approximable and moreover its continued fraction expansion contains all patterns.*

We would like to point out that the special invariance properties of the Cantor set are actually crucial for the proof of Theroem 3.1 while the result concerning the intersection of the Cantor set with BA are much more general. The same method that gives Theorem 3.1 can also be used for $d = 2$ together with Theorem 1.4 and leads to the following theorem.

Theorem 3.2. *Let $A : \mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ be a hyperbolic automorphism, induced by the linear action of a matrix $A \in \mathrm{SL}_2(\mathbb{Z})$ and let μ be a probability measure which is invariant and ergodic with respect to A, and has positive dimension. Then μ almost any $v \in \mathbb{R}^2/\mathbb{Z}^2$ is well approximable.*

To see the connection between these two theorems and Theorem 2.2 resp. Theorem 1.4 we need to recall the interpretation of $X_n = \mathrm{SL}_n(\mathbb{Z})\backslash \mathrm{SL}_n(\mathbb{R})$ as the space of unimodular lattices in $\mathbb{R}^n$. In fact, we may identify the identity coset $\mathrm{SL}_n(\mathbb{Z})$ with the unimodular (i.e. covolume one) lattice $\Lambda = \mathbb{Z}^n$. More generally, we identify the coset $\mathrm{SL}_n(\mathbb{Z})g$ with the lattice $\Lambda = \mathbb{Z}^n g$. This gives an isomorphism between X_n and the space of unimodular lattices in $\mathbb{R}^n$, and makes it possible to classify compact subsets by the following geometric property.

Theorem 3.3 (Mahler's compactness criterion). *A subset $C \subset X_n$ is bounded (i.e. its closure is compact) if and only if there exists $\epsilon > 0$ such that for any lattice $\Lambda \in C$, $\Lambda \cap B_\epsilon(0) = \emptyset$ i.e. if and only if there exists a uniform lower bound for the lengths of nonzero vectors belongings to points in C.*

This gives the basis of the dynamical interpretation of badly approximable vectors $\mathbf{v}$ (used as row vector) in terms of the orbit of the associated lattice

$$\Lambda_{\mathbf{v}} = \mathbb{Z}^{d+1} \begin{pmatrix} I_d & 0 \\ \mathbf{v} & 1 \end{pmatrix}$$

with respect to the generalization of the geodesic flow defined below. Here we write I_d for the $d \times d$-identity matrix.

Corollary 3.4. *We define the diagonal elements*

$$a_t = \begin{pmatrix} e^{t/d} I_d & 0 \\ 0 & e^{-t} \end{pmatrix}$$

for any $t \in \mathbb{R}$. Then a vector $\mathbf{v} \in \mathbb{R}^d$ is badly approximable if and only if the forward orbit

$$\{\Lambda_{\mathbf{v}} a_t : t \geq 0\}$$

of the lattice $\Lambda_{\mathbf{v}}$ associated to $\mathbf{v}$ is bounded in X_{d+1}.

Let us indicate one direction of this characterization. If $\mathbf{v}$ is badly approximable as in (1) and $t \geq 0$, then the elements of the lattice $\Lambda_{\mathbf{v}} a_t$ are of the form

$$((\mathbf{p} - q\mathbf{v})e^{t/d}, e^{-t}q).$$

We claim that any non-zero such element cannot be closer to the origin in $\mathbb{R}^{d+1}$ than c. Otherwise, we derive from $t \geq 0$ that $q \neq 0$ and by taking the product of the norm of $(\mathbf{p} - q\mathbf{v})e^{t/d}$ and of the d-th root of $e^{-t}q$ that $q^{1/d}\|\mathbf{p} - q\mathbf{v}\| < c$ — a contradiction to (1). The opposite implication is similar.

Let us indicate now the relationship between Theorem 2.2 and Theorem 3.1. Write ν_C for the uniform measure on the middle third Cantor set. We may embed ν_C as a measure on X_2 by push-forward via the map $v \to \Lambda_v$. By Corollary 3.4 what we would like to show is that for ν_C-a.e. point the orbit under the geodesic flow is unbounded.

To better phrase the special invariance properties that the Cantor set has, it makes sense to introduce the 3-adic extension of X_2. One can check that

$$X_2 \simeq \mathrm{SL}_2\left(\mathbb{Z}\left[\frac{1}{3}\right]\right) \backslash \mathrm{SL}_2(\mathbb{R}) \times \mathrm{SL}_2(\mathbb{Q}_3) / \mathrm{SL}_2(\mathbb{Z}_3),$$

so that we should think of $\widetilde{X_2} = \mathrm{SL}_2(\mathbb{Z}[\frac{1}{3}]) \backslash \mathrm{SL}_2(\mathbb{R}) \times \mathrm{SL}_2(\mathbb{Q}_3)$ as an extension of X_2 by compact fibers isomorphic to $\mathrm{SL}_2(\mathbb{Z}_3)$.

We note that

$$\mathrm{SL}_2(\mathbb{Z}[\tfrac{1}{3}])\begin{pmatrix}1&0\\v&1\end{pmatrix}\begin{pmatrix}3&0\\0&3^{-1}\end{pmatrix}=\mathrm{SL}_2\left(\mathbb{Z}\left[\frac{1}{3}\right]\right)\begin{pmatrix}1&0\\9v&1\end{pmatrix},$$

which shows that right-multiplication by the diagonal element $\begin{pmatrix}3&0\\0&3^{-1}\end{pmatrix}$ corresponds to multipying v by 9. As the Cantor set has a special relationship with respect to multiplication by 3 (or equivalently ternary digit expansions), this can be exploited and one can construct an invariant measure $\widetilde{\nu_C}$ on $\widetilde{X_2}$ for the map that multiplies on the right — both in the real and 3-adic component — by $\begin{pmatrix}3&0\\0&3^{-1}\end{pmatrix}$. However, this dynamical system is different from the extension of the geodesic flow a_t, which is just right multiplication by the diagonal element $\begin{pmatrix}e^t&0\\0&e^{-t}\end{pmatrix}$ in the real component. Taking the average of $\widetilde{\nu_C}$ along the orbit under a_t one obtains a measure that is invariant under the diagonal subgroup — both in the real component and the 3-adic component. To this limit measure μ one can apply Theorem 2.2. The recurrence condition is assured since μ is actually invariant under a non-compact subgroup of $\mathrm{SL}_2(\mathbb{Q}_3)$. The entropy assumption is satisfied in the weaker sense that the entropy of μ is positive — this is a consequence of the fact that ν_C had positive dimension. From this, we conclude not necessarily that μ equals the Haar measure but at least that it contains the Haar measure as one of the ergodic components. Clearly the Haar measure has non-compact support, and this can be used to deduce[2] Theorem 3.1.

4. Non-uniformity of Bad Approximations

Recall that every irrational $x\in[0,1]$ can be written as a continued fraction. The digits of the continued fraction expansion relates to the discussion of the Diophantine approximation above. In fact, x is badly approximable if and only if the digits $a_n(x)$ of the expansion are bounded. If x is badly approximable, then the quantity $c(x)=\limsup a_n(x)$ measures the extend to which x is badly approximable. In this sense, the next theorem says that the sequence $x,2x,\dots,nx,\dots$ cannot be uniformly badly approximable.

Theorem 4.1. *If we denote for $v\in[0,1]$, $c(v)=\limsup a_n(v)$ where $a_n(v)$ are the coefficients in the continued fraction expansion of v, then for any irrational $v\in[0,1]$, $\sup_n c(n^2v)=\infty$, where n^2v is calculated modulo 1.*

[2]The cautious reader may notice that what we said implies only that the quantity c as in (1) cannot be uniform for a.e. point in C, but with a bit more work, using only ergodicity of the Haar measure, one really obtains a proof.

This is also joint work with L. Fishman and U. Shapira [10]. We would like to point out that this relates to a conjecture of M. Boshernitzan, who reported to us that a stronger version of Theorem 4.1 holds for the special case of quadratic irrationals.

The proof of Theorem 4.1 is similar in spirit to the proof of Theorem 3.1, but this time takes place on

$$X_{2,\mathbb{A}} = \mathrm{SL}_2(\mathbb{Q})\backslash \mathrm{SL}_2(\mathbb{A})$$

where multiplication of v by n^2 can be converted to right multiplication by the matrix $\begin{pmatrix} n & 0 \\ 0 & n^{-1} \end{pmatrix}$ (in every component). These elements together with the real diagonal subgroup give a big subgroup A' of the full group $A_{\mathbb{A}}$ of adelic points of the diagonal subgroup, more precisely the quotient of $A_{\mathbb{A}}$ by A' is compact. In this theorem there is no mention of entropy or dimension due to the following theorem by E. Lindenstrauss [35] (which is the combination of Theorem 2.2 and the method in [2]).

Theorem 4.2. *The action of the group, $A_{\mathbb{A}}$, of adelic points of the diagonal subgroup in SL_2 on $X_{2,\mathbb{A}} = \mathrm{SL}_2(\mathbb{Q})\backslash \mathrm{SL}_2(\mathbb{A})$ is uniquely ergodic.*

5. Littlewood's Conjecture

Historically the second application of a partial measure classification result for diagonal subgroups (in this case Theorem 1.4) has been a partial result towards Littlewood's conjecture.

Conjecture 5.1 (Littlewood (c. 1930)). *For every $\alpha, \beta \in \mathbb{R}$,*

$$\liminf_{n\to\infty} n\|n\alpha\|\|n\beta\| = 0, \tag{2}$$

where $\|w\| = \min_{n\in\mathbb{Z}} |w - n|$ is the distance of $w \in \mathbb{R}$ to the nearest integer.

Similar to Corollary 3.4 one can also show the following characterization of Littlewood's conjecture in dynamical terms.

Proposition 5.2. *(α, β) satisfy (2) if and only if the orbit*

$$\Lambda_{\alpha,\beta} a_{s,t} = \mathrm{SL}(3,\mathbb{Z}) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \alpha & \beta & 1 \end{pmatrix} a_{s,t}$$

under the semigroup

$$A^+ = \{a(s,t) : s,t \geq 0\} \qquad a(s,t) = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^s & 0 \\ 0 & 0 & e^{-t-s} \end{pmatrix}$$

is unbounded in $X_3 = \mathrm{SL}(3,\mathbb{Z})\backslash \mathrm{SL}(3,\mathbb{R})$.

Together with Theorem 1.4 this leads to the following theorem.

Theorem 5.3 ([12, Theorem 1.5]). *For any $\delta > 0$, the set*[3]

$$\Xi_\delta = \left\{(\alpha,\beta) \in [0,1]^2 : \liminf_{n\to\infty} n\|n\alpha\|\|n\beta\| \geq \delta\right\}$$

has zero upper box dimension[4]*. In particular, $\bigcup_{\delta>0}\Xi_\delta$ has zero Hausdorff dimension.*

We refer to [12] or to [15] for an explanation on how the entropy assumption in Theorem 1.4 is converted to the box dimension result above. A full solution of either Conjecture 1.1 or Conjecture 1.2 would imply Conjecture 5.1.

The same method can also be used to obtain a partial result towards a conjecture of B. de Mathan and O. Teulié [5]. They conjectured[5] that for every prime number p, for every $u \in \mathbb{R}$ and $\epsilon > 0$

$$\left|qu - q_0\right| < \frac{\epsilon}{q|q|_p} = \frac{\epsilon}{q'} \quad \text{for infinitely many pairs } (q, q_0) \in \mathbb{Z}^2,$$

where $q = q'p^k$ for some $k \geq 0$, q' is coprime to p, and $|q|_p = 1/p^k$ denotes the p-adic norm. Equivalently one can ask whether

$$\liminf_{q\to\infty} |q| \cdot |q|_p \cdot \|qu\| = 0\,, \tag{3}$$

In joint work with Kleinbock [13] we have shown the following analogue to Theorem 5.3.

Theorem 5.4. *The set of $u \in \mathbb{R}$ which do not satisfy* (3) *has Hausdorff dimension zero.*

6. Compact Orbits and Ideal Classes

Another interesting connection between the dynamics of the full diagonal subgroup A on $X_n = \mathrm{SL}_n(\mathbb{Z})\backslash \mathrm{SL}_n(\mathbb{R})$ and number theory arises in the study of periodic (i.e. compact) orbits of A on X_n.

In fact, if $I \subset \mathcal{O}_K$ is an ideal in the ring of integers of a totally real number field K of degree n then this ideal can give rise to a compact A-orbit as follows. To see this, let $\phi_1, \dots, \phi_n : K \to \mathbb{R}$ be the complete list of Galois embeddings. Then

$$\{(\phi_1(k), \dots, \phi_n(k)) : k \in I\} \subset \mathbb{R}^n$$

[3]Since (2) depends only on α, β mod 1 it is sufficient to consider only $(\alpha,\beta) \in [0,1]^2$.

[4]I.e., for every $\epsilon > 0$, for every $0 < r < 1$, one can cover Ξ_δ by $O_{\delta,\epsilon}(r^{-\epsilon})$ boxes of size $r \times r$.

[5]Their conjecture is more general.

is a lattice in $\mathbb{R}^n$, which after normalization of the covolume, gives an element $\Lambda_J \in X_n$. By Dirichlet's unit theorem, there are $n-1$ multiplicatively independent units in the ring $\mathcal{O}_K$. Let ξ be one such unit. Replacing ξ by ξ^2 if necessary, we may assume that $\phi_i(\xi) > 0$ for all i. Then $a = \mathrm{diag}(\phi_1(\xi), \ldots, \phi_n(\xi)) \in A$ satisfies

$$\{(\phi_1(k), \ldots, \phi_n(k)) : k \in I\}a = \{(\phi_1(k), \ldots, \phi_n(k)) : k \in \xi I\},$$

which shows that $\Lambda_I = \Lambda_I a$ is fixed under a since $\xi I = I$ for any unit. As A has $n-1$ dimensions and we have $n-1$ independent units, one obtains that $\Lambda_I A$ is an $n-1$-dimensional torus and so compact. We write $\mu_{\Lambda_I A}$ for the Lebesgue measure on this torus normalized to be a probability measure and viewed as a meaure on X_n.

One can furthermore check that two ideals give rise to the same compact orbit if and only if the two ideals are equivalent. Therefore, the number of compact A orbits arising from the maximal order $\mathcal{O}_K$ of the field is precisely the class number of the field.

The same construction shows that any ideal in any order $\mathcal{O}$ of K gives rise to a compact A-orbit. Allowing this more general construction one actually obtains all compact A-orbits. It is natural to ask how the various compact orbits for a given order distribute within X_n. If $n = 2$ special cases of the expected equidistribution theorem have been proven around 1960 by Linnik [37] and Skubenko [54]. The full statement has been proven by Duke [7] in 1988 using subconvexity estimates of L-functions. For $n = 3$ the analogue has been obtained more recently in joint work with Lindenstrauss, Michel, and Venkatesh [9].

Theorem 6.1. *Let K_ℓ be a sequence of totally real degree three extensions of $\mathbb{Q}$, and let h_ℓ the class number of K_ℓ. Let $x_{1,\ell}A, \ldots, x_{h_\ell,\ell}A \subset X_3$ be the periodic A orbits corresponding to the ideal classes of K_ℓ as above. Let $\mu_\ell = \frac{1}{h_\ell}\sum_i \mu_{x_{i,\ell}A}$. Then μ_ℓ converge in the weak* topology to the* $\mathrm{SL}(3,\mathbb{R})$ *invariant probability measure m_{X_3} on X_3.*

The proof uses a combination of methods. First, subconvexity estimates of Duke, Friedlander and Iwaniec [8] imply that for certain test functions f, the integrals $\int_{X_3} f d\mu_\ell$ converge to the expected value (i.e. $\int_{X_3} f dm_{X_3}$). The space of these test functions is not sufficient to conclude Theorem 6.1, but can be used to deduce that μ_ℓ is a probability measure (i.e. there is no escape of mass to the cusp) and that the entropy of *every* ergodic component in such a limiting measure is greater than an explicit lower bound. Once these two facts have been established, Theorem 1.4 gives the result.

We also refer to [9, 16] and the survey [39] for more details on this and related applications.

7. Counting Rational Points

For the following application due to Zamojski we fix a monic irreducible polynomial $P(\lambda) \in \mathbb{Q}[\lambda]$ of degree n. Let us assume that $P(\lambda)$ factorizes over $\mathbb{R}$. Let $V \subset \mathrm{Mat}_{nn}$ be the variety consisting of all matrices whose characteristic polynomial equals $P(\lambda)$. Next recall that for any rational vector $v = (\frac{p_1}{q}, \dots, \frac{p_\ell}{q})$ represented in lowest terms, we can define the *height* as the maximum of the absolute values $|p_i|$ and the common denominator q. Zamojski [60] has proven the following asymptotic counting formula.

Theorem 7.1. *If N_R denotes the number of rational matrices with characteristic polynomial $P(\lambda)$ and height bounded by R, then the limit*

$$\lim_{R\to\infty} \frac{N_R}{R^{n(n-1)/2+1}}$$

exists and is positive.

This proves a new case of Manin's conjecture (see [57, 58]) which concerns similar counting problems on more general varieties. There is already a rich history for the interaction between asymptotic counting problems and ergodic theory. Initially, only mixing in the form of the theorem by Howe and Moore was used, see for instance the influential work of Eskin and McMullen [19]. However, after Ratner proved her theorems [44] further cases of the counting problem could be handled. For instance, Eskin, Mozes, and Shah [20] have proven in 1996 the integer version of Theorem 7.1.

In all of these proofs of asymptotic counting the following equidistribution problem is of crucial importance. The variety V as above is actually a single orbit of $\mathrm{PGL}_n(\mathbb{R})$, we write H for its stabilizer. Similar to the discussion in §6 the orbit $\mathrm{PGL}_n(\mathbb{Z})H$ of the identity coset is compact, we write μ for the measure on $X = \mathrm{PGL}_n(\mathbb{Z})\backslash \mathrm{PGL}_n(\mathbb{R})$ that is supported on $\mathrm{PGL}_n(\mathbb{Z})H$ and invariant under H. Then the counting problem of integer points in V is related to the equidistribution of the measure μg (obtained by applying right multiplication by g) on the space X. In [20] it is shown that, at least on average, μg indeed equidistributes. For this the theory of unipotent dynamics was used, which at first may be surprising as H does not contain any unipotents. The key link of this problem to unipotents lies in the fact that μg is invariant under $g^{-1}Hg$, which if $g_n \to \infty$ in $H\backslash G$ implies that any limit measure of μg_n is invariant under a one-parameter unipotent subgroup.

For counting rational points on homogenous varieties it is natural to replace the quotient X by the corresponding adelic quotient, as was shown in the work of Gorodnik, Maucourant, and Oh [22]. Also in the proof of Theorem 7.1 the equidistribution of translates of a given finite volume measure on the adelic quotient $\mathrm{PGL}_n(\mathbb{Q})\backslash \mathrm{PGL}_n(\mathbb{A})$ is studied. However, unlike the case of counting integer points, it is no longer true that the translated measures will on average develop invariance properties under a unipotent subgroup. Roughly speaking

this is because it is not true that if a sequence $g_n \in \mathrm{PGL}_n(\mathbb{A})$ goes to infinity, then for some place p the projection of g_n to this place goes to infinity. Indeed, as Zamojski shows for most sequences g_n the projections stay bounded within each place. Hence the limit measures are only known to have the same invariance as the original measure μ — invariance under a conjugate of the diagonal subgroup A. Zamojski shows, similar to a part of the proof of Theorem 6.1 in [9] that a limit measure must have positive entropy (for all of its ergodic components) and so Theorem 1.4 can be applied. However, if n is not a prime number, Zamojski gives an additional argument which rules out the measures corresponding to intermediate subgroups.

8. Divisibility Properties of Hamiltonian Quaternions

Our last application uses an analogue of Theorem 2.2 for a quotient of the form $\Gamma \backslash \mathrm{PGL}_2(\mathbb{Q}_p) \times \mathrm{PGL}_2(\mathbb{Q}_q)$ for two primes $p \neq q$, and concerns divisibility properties of integer Hamiltonian quaternions. This is a special case of ongoing joint work with S. Mozes.

Let $H = \mathbb{R}[i, j, k]$ be the Hamiltonian quaternions, and let $\mathcal{O} = \mathbb{Z}[i, j, k]$ be the order consisting of integer combinations of 1 and the three imaginary units i, j, k. We write $N(a + bi + cj + dk) = a^2 + b^2 + c^2 + d^2$ for the norm on H.

Let $p \neq 2$ be a prime number. Then

$$\Gamma_p = \{\alpha \in \mathcal{O} : N(\alpha) \text{ is a power of } p\}$$

is a multiplicative semi-group, for which ± 1 and p generates the center C. Taking the quotient by the center, one obtains a group $\mathrm{P}\,\Gamma_p = \Gamma_p / C$. As a consequence of Pall's unique factorization theorem for elements of $\mathcal{O}$ it follows that $\mathrm{P}\,\Gamma_p$ is virtually a free group (more concretely it contains a free group with $\frac{p+1}{2}$ generators and index 4).

Similarly if $p, q \neq 2$ are two different odd prime numbers, then we define the semigroup

$$\Gamma_{p,q} = \{\alpha \in \mathcal{O} : N(\alpha) \text{ is a product of powers of } p \text{ and } q\}$$

which once more gives a group $\mathrm{P}\,\Gamma_{p,q} = \Gamma_{p,q}/C$ after dividing by the center.

The group $\mathrm{P}\,\Gamma_{p,q}$ is far from being a free group. This is known, but is also shown clearly by the following theorem. For stating the theorem we need some definitions. We say an element $\alpha \in \mathcal{O}$ appears[6] in $\beta \in \mathcal{O}$ if there exits some $\ell, r \in \mathcal{O}$ such that $\beta = \ell \alpha r$. The fact that $\mathrm{P}\,\Gamma_p$ contains a free subgroup shows that for any fixed $\alpha \in \Gamma_p$ with $N(\alpha) > 1$ the set

$$\{\beta \in \Gamma_p : \alpha \text{ does not appear in } \beta \text{ and } N(\beta) = p^k\}$$

[6]We write "appears" for this notion of divisibility to distinguish this notion from a left- or right-divisibility that is sometimes considered for non-commutative rings.

grows exponentially with k. We say that $\alpha \in \Gamma_{p,q}$ is reduced if $\frac{1}{p}\alpha \notin \mathcal{O}$ and $\frac{1}{q}\alpha \notin \mathcal{O}$. In contrast to Γ_p we have the following theorem concerning $\Gamma_{p,q}$.

Theorem 8.1. *Let $p, q \neq 2$ be two different primes. Then for any reduced $\alpha \in \Gamma_{p,q}$ the set*

$$\{\beta \in \Gamma_{p,q} : \alpha \text{ does not appear in } \beta \text{ and } N(\beta) = p^k q^k\}$$

grows sub-exponentially. That is, if $M(k)$ is the cardinality of the set, then

$$\lim_{k\to\infty} \frac{1}{k} \log M(k) = 0.$$

Let us indicate the connection between the above theorem and the dynamics of diagonal flows, which goes back to [41]. First we may choose a subgroup $\Gamma \subset \mathrm{P}\,\Gamma_{p,q}$ of finite index which does not contain the images of the elements $\pm i, \pm j, \pm k$ and is generated by two free subgroups of $\mathrm{P}\,\Gamma_p$ and $\mathrm{P}\,\Gamma_q$. Then Γ is naturally a lattice in the group $G = \mathrm{PGL}_2(\mathbb{Q}_p) \times \mathrm{PGL}_2(\mathbb{Q}_q)$ for which a fundamental domain is given by the compact set $F = \mathrm{PGL}_2(\mathbb{Z}_p) \times \mathrm{PGL}_2(\mathbb{Z}_q)$.

Let $a_p = \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \times I$, where I denotes the identity. Define a_q similarly. Since F is a fundamental domain, there exists for every $f \in F$ some element $\gamma \in \Gamma \cap \mathrm{P}\,\Gamma_p$ and some $f' \in F$ for which $fa_p = \gamma f'$. Clearly, if f is replaced by a slight perturbation of f, then γ will not change. In this sense, γ corresponds to an open subset O_γ of $F \simeq \Gamma\backslash G$. More generally, if $\gamma \in \Gamma$ then γ is the image of some reduced element in $\mathcal{O}$ which we assume has norm $p^k q^k$ and we can define the open (and non-empty) subset

$$O_\gamma = \{f \in F : fa_p^k a_q^k \in \gamma F\}.$$

If now $\beta = l\alpha r$ for elements $\alpha, \beta, l, r \in \mathcal{O}$ with $N(l) = p^m q^n$, then $f \in O_\beta$ implies that $fa_p^m a_q^n \in O_\alpha$. This has a partial converse, meaning that if $f \in O_\beta$ satisfies $fa_p^m a_q^n \in \Gamma O_\alpha$ for sufficiently small values of m and n then we deduce that α appears in β.

In this sense an element β of norm $N(\beta) = p^k q^k$ in which α does not appear, gives rise to a piece of an orbit under the joint action of a_p and a_q on $\Gamma\backslash G$ that does not visit the open set ΓO_α. If there are exponentially many such elements β as $k \to \infty$, then one can construct (with the help of the variational principle from ergodic theory) from these many large pieces of orbits an invariant measure on $\Gamma\backslash G$ with positive entropy and zero mass on the set O_α. The analogue of Theorem 1.4 for the action of a_p and a_q on $\Gamma\backslash G$ holds and is indeed a version of Theorem 2.2, hence we derive a contradiction and Theorem 8.1 follows.

9. Open Problems

We already mentioned the main open problems: Furstenberg's Conjecture 1.3 regarding jointly invariant probability measures for the times 2 and times 3

maps on $\mathbb{T}$, and Margulis' Conjectures 1.1–1.2 regarding bounded orbits and invariant measures on $\mathrm{SL}_n(\mathbb{Z})\backslash \mathrm{SL}_n(\mathbb{R})$ for $n \geq 3$. We also refer to [14] and [23] for related conjectures on the measure classification.

However, even if we allow ourselves the positive entropy assumption there are still unsolved cases where no analogue to Theorem 1.4 is known. For instance we may take a subgroup $A' \subset A$ of dimension two within the three-dimensional diagonal subgroup $A \subset \mathrm{SL}_4(\mathbb{R})$ and ask what are the A'-invariant and ergodic probability measures on $X_4 = \mathrm{SL}_4(\mathbb{Z})\backslash \mathrm{SL}_4(\mathbb{R})$ for which some element $a \in A'$ acts with positive entropy. The current techniques that go into Theorem 1.4 fall short in this case.

The list of the applications, discussed above, also suggests a number of open problems. For instance, Theorem 3.2 currently only holds for $d = 2$ and Theorem 4.1 only for $d = 1$. However, we certainly would expect that these hold in any dimension.

Also Theorem 6.1 currently only holds for cubic fields and the non-compact space X_3, so it is natural to ask for the same for higher dimensions or for compact quotients of $\mathrm{SL}_3(\mathbb{R})$ by the units in a degree 3 division algebra over $\mathbb{Q}$.

Another interesting question arises by comparing the argument in [10] (see §3) with Host's theorem [24, 25].

Conjecture 9.1. *Let μ be a probability measure on an irreducible quotient $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R}) \times \mathrm{SL}_2(\mathbb{R})$. Suppose μ is invariant and ergodic with respect to the action of the one-parameter diagonal subgroup $A_1 \subset \mathrm{SL}_2(\mathbb{R}) \times \{1\}$ of the first copy of $\mathrm{SL}_2(\mathbb{R})$, and suppose μ has positive entropy with respect to A_1. Write $A_2 \subset \{1\} \times \mathrm{SL}_2(\mathbb{R})$ for the one-parameter diagonal subgroup in the second $\mathrm{SL}_2(\mathbb{R})$. Then μ-a.e. $x \in X$ has equidistributed orbit for the action of A_2.*

The theorem in [24] concerns the same problem with $X = \mathbb{T}$, A_1 replaced by $\times 2$, and A_2 replaced by $\times 3$. A slightly easier problem would be to generalize the related theorem of Johnson and Rudolph [26], which might look as follows.

Conjecture 9.2. *Let μ be a probability measure on an irreducible quotient $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R}) \times \mathrm{SL}_2(\mathbb{R})$. Suppose μ is invariant and ergodic with respect to the action of the one-parameter diagonal subgroup $A_1 \subset \mathrm{SL}_2(\mathbb{R}) \times \{1\}$ of the first copy of $\mathrm{SL}_2(\mathbb{R})$, and suppose μ has positive entropy with respect to A_1. Write $A_2 \subset \{1\} \times \mathrm{SL}_2(\mathbb{R})$ for the one-parameter diagonal subgroup in the second $\mathrm{SL}_2(\mathbb{R})$. Then*

$$\frac{1}{T} \int_X \int_0^T f(x a_{2,t}) \, \mathrm{d}t \, \mathrm{d}\mu(x).$$

converges to $\int f \, \mathrm{d}m_X$ where $f \in C_c(X)$ and $a_{2,t} \in A_2$ denotes a homomorphism from $\mathbb{R}$ to A_2.

Similarly, the above two conjectures can be asked for other quotients for which the analogue of Theorem 2.2 or Theorem 1.4 holds.

References

[1] E Artin, *Ein mechanisches System mit quasiergodischen Bahnen*, Hamb. Math. Abh. **3** (1924), 170–177.

[2] Jean Bourgain and Elon Lindenstrauss, *Entropy of quantum limits*, Comm. Math. Phys. **233** (2003), no. 1, 153–171.

[3] S. G Dani, *Divergent trajectories of flows on homogeneous spaces and diophantine approximation*, J. Reine Angew. Math. **359** (1985), 55–89.

[4] S. G Dani and John Smillie, *Uniform distribution of horocycle orbits for Fuchsian groups*, Duke Math. J. **51** (1984), no. 1, 185–194.

[5] Bernard de Mathan and Olivier Teulié, *Problèmes diophantiens simultanés*, Monatsh. Math. **143** (2004), no. 3, 229–245.

[6] Y Colin de Verdière, *Ergodicité et fonctions propres du laplacien*, Comm. Math. Phys. **102** (1985), no. 3, 497–502.

[7] W Duke, *Hyperbolic distribution problems and half-integral weight maass forms*, Invent. Math. **92** (1988), no. 1, 73–90.

[8] W Duke, J B Friedlander, and H Iwaniec, *The subconvexity problem for Artin L-functions*, Invent. Math. (2002), no. 149 (3), 489–577.

[9] M Einsiedler, E Lindenstrauss, Ph Michel, and A Venkatesh, *Distribution of periodic torus orbits and Duke's theorem for cubic fields*, to appear in Ann. Math.

[10] Manfred Einsiedler, Lior Fishman, and Uri Shapira, *Diophantine approximations on fractals*, arXiv **math.DS** (2009), 20 pages.

[11] Manfred Einsiedler and Anatole Katok, *Invariant measures on G/Gamma for split simple Lie groups G*, Comm. Pure Appl. Math. **56** (2003), no. 8, 1184–1221.

[12] Manfred Einsiedler, Anatole Katok, and Elon Lindenstrauss, *Invariant measures and the set of exceptions to Littlewood's conjecture*, Ann. of Math. (2) **164** (2006), no. 2, 513–560.

[13] Manfred Einsiedler and Dmitry Kleinbock, *Measure rigidity and p-adic littlewood-type problems*, Compos. Math. **143** (2007), no. 3, 689–702.

[14] Manfred Einsiedler and Elon Lindenstrauss, *Diagonalizable flows on locally homogeneous spaces and number theory*, International Congress of Mathematicians. Vol. II, 1731–1759, Eur. Math. Soc., Zürich, 2006. (2006), 1731–1759.

[15] ———, *Diagonal actions on locally homogeneous spaces*, to appear in Clay Mathematics Proceedings (2010).

[16] Manfred Einsiedler, Elon Lindenstrauss, Philippe Michel, and Akshay Venkatesh, *Distribution of periodic torus orbits on homogeneous spaces*, Duke Math. J. **148** (2009), no. 1, 119–174.

[17] Manfred Einsiedler and Thomas Ward, *Arithmetic quantum unique ergodicity on* $\Gamma\backslash\mathbb{H}$, Arizona Winter School 2010: Number theory and Dynamics.

[18] ———, *Ergodic theory with a view towards number theory*, to appear in Springer GTM (2010), 1–480.

[19] Alex Eskin and Curt McMullen, *Mixing, counting, and equidistribution in lie groups*, Duke Math. J. **71** (1993), no. 1, 181–209.

[20] Alex Eskin, Shahar Mozes, and Nimish Shah, *Unipotent flows and counting lattice points on homogeneous varieties*, Ann. of Math. (2) **143** (1996), no. 2, 253–299.

[21] Harry Furstenberg, *Disjointness in ergodic theory, minimal sets, and a problem in diophantine approximation*, Math. Systems Theory **1** (1967), 1–49.

[22] Alex Gorodnik, Francois Maucourant, and Hee Oh, *Manin's and peyre's conjectures on rational points and adelic mixing*, Ann. Sci. Ec. Norm. Supér. (4) **41** (2008), no. 3, 383–435.

[23] Alexander Gorodnik, *Open problems in dynamics and related fields*, J. Mod. Dyn. **1** (2007), no. 1, 1–35.

[24] Bernard Host, *Nombres normaux, entropie, translations*, Israel J. Math. **91** (1995), no. 1-3, 419–428.

[25] ———, *Some results of uniform distribution in the multidimensional torus*, Ergodic Theory Dynam. Systems **20** (2000), no. 2, 439–452.

[26] Aimee Johnson and Daniel J Rudolph, *Convergence under $\times_q$ of $\times_p$ invariant measures on the circle*, Adv. Math. **115** (1995), no. 1, 117–140.

[27] Boris Kalinin and Anatole Katok, *Invariant measures for actions of higher rank abelian groups*, **69** (2001), 593–637.

[28] Boris Kalinin and Ralf Spatzier, *Rigidity of the measurable structure for algebraic actions of higher-rank abelian groups*, Ergodic Theory Dynam. Systems **25** (2005), no. 1, 175–200.

[29] A Katok and R. J Spatzier, *Invariant measures for higher-rank hyperbolic abelian actions*, Ergodic Theory Dynam. Systems **16** (1996), no. 4, 751–778.

[30] ———, *Corrections to: "Invariant measures for higher-rank hyperbolic abelian actions" [Ergodic Theory Dynam. Systems 16 (1996), no. 4, 751-778]*, Ergodic Theory Dynam. Systems **18** (1998), no. 2, 503–507.

[31] Anatole Katok and Boris Hasselblatt, *Introduction to the modern theory of dynamical systems*, Cambridge University Press, Cambridge **54** (1995), xviii+802.

[32] Dmitry Kleinbock, Nimish Shah, and Alexander Starkov, *Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory*, Handbook of dynamical systems, Vol. 1A (2002), no. North-Holland, Amsterdam, 813–930.

[33] Dmitry Kleinbock and Barak Weiss, *Badly approximable vectors on fractals*, Israel J. Math. **149** (2005), 137–170.

[34] Simon Kristensen, Rebecca Thorn, and Sanju Velani, *Diophantine approximation and badly approximable sets*, Adv. Math. **203** (2006), no. 1, 132–169.

[35] Elon Lindenstrauss, *Adelic dynamics and arithmetic quantum unique ergodicity*, Current developments in mathematics, 2004, 111–139, Int. Press, Somerville (2006), 111–139.

[36] ———, *Invariant measures and arithmetic quantum unique ergodicity*, Ann. of Math. (2) **163** (2006), no. 1, 165–219.

[37] Yu V Linnik, *Ergodic properties of algebraic fields*, Translated from the Russian by M. S. Keane. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 45. Springer-Verlag New York Inc., New York, 1968.

[38] G. A Margulis, *Indefinite quadratic forms and unipotent flows on homogeneous spaces*, Dynamical systems and ergodic theory (Warsaw, 1986) **23** (1989), 399–409.

[39] Philippe Michel and Akshay Venkatesh, *Equidistribution, L-functions and ergodic theory: on some problems of Yu. Linnik*, International Congress of Mathematicians. Vol. II, 421–457, Eur. Math. Soc., Zürich, 2006. (2006).

[40] Dave Witte Morris, *Ratner's theorems on unipotent flows*, University of Chicago Press, Chicago, IL (2005), xii+203.

[41] Shahar Mozes, *A zero entropy, mixing of all orders tiling system*, **135** (1992), 319–325.

[42] Marina Ratner, *Factors of horocycle flows*, Ergodic Theory Dynam. Systems **2** (1982), no. 3–4, 465–489 (1983).

[43] ———, *Horocycle flows, joinings and rigidity of products*, Ann. of Math. (2) **118** (1983), no. 2, 277–313.

[44] ———, *On Raghunathan's measure conjecture*, Ann. of Math. (2) **134** (1991), no. 3, 545–607.

[45] ———, *Raghunathan's topological conjecture and distributions of unipotent flows*, Duke Math. J. **63** (1991), no. 1, 235–280.

[46] ———, *Interactions between ergodic theory, Lie groups, and number theory*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994) (1995), 157–182.

[47] Zeév Rudnick and Peter Sarnak, *The behaviour of eigenstates of arithmetic hyperbolic manifolds*, Comm. Math. Phys. **161** (1994), no. 1, 195–213.

[48] Daniel J Rudolph, $\times 2$ *and* $\times 3$ *invariant measures and entropy*, Ergodic Theory Dynam. Systems **10** (1990), no. 2, 395–406.

[49] Peter Sarnak, *Recent progress on QUE*.

[50] ———, *Spectra of hyperbolic surfaces*, Bull. Amer. Math. Soc. (N.S.) **40** (2003), no. 4, 441–478 (electronic).

[51] Caroline Series, *The modular surface and continued fractions*, J. London Math. Soc. (2) **31** (1985), no. 1, 69–80.

[52] ———, *Geometrical methods of symbolic coding*, (1991), 125–151.

[53] Lior Silberman and Akshay Venkatesh, *On quantum unique ergodicity for locally symmetric spaces*, Geom. Funct. Anal. **17** (2007), no. 3, 960–998.

[54] B F Skubenko, *The asymptotic distribution of integers on a hyperboloid of one sheet and ergodic theorems*, Izv. Akad. Nauk SSSR Ser. Mat. (1962), 26:721—752.

[55] A. I Snirelprimeman, *Ergodic properties of eigenfunctions*, Uspekhi Mat. Nauk **29** (1974), no. 6(180), 181–182.

[56] K Soundararajan, *Quantum unique ergodicity for* $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, Ann. of Math. (2).

[57] Yuri Tschinkel, *Fujita's program and rational points*, Higher dimensional varieties and rational points (Budapest, 2001), 283–310, Bolyai Soc. Math. Stud., 12, Springer, Berlin, 2003.

[58] ———, *Geometry over nonclosed fields*, International Congress of Mathematicians. Vol. II, 637–651, Eur. Math. Soc., Zürich, 2006.

[59] Thomas C Watson, *Rankin triple products and quantum chaos*, arXiv **math.NT** (2008).

[60] Thomas Zamojski, *Counting rational matrices of a fixed irreducible characteristic polynomial*, in preparation (2010).

[61] Steven Zelditch, *Uniform distribution of eigenfunctions on compact hyperbolic surfaces*, Duke Math. J. **55** (1987), no. 4, 919–941.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Measure Theory and Geometric Topology in Dynamics

Federico Rodriguez Hertz*

Abstract

In this survey we shall present some relations between measure theory and geometric topology in dynamics. One of these relations comes as follows, on one hand from topological information of the system, some structure should be preserved by the dynamics at least in some weak sense, on the other hand, measure theory is soft enough that an invariant geometric structure almost always appears along some carefully chosen invariant measure. As an example, we have the known result that in dimension 2 the system has asymptotic growth of hyperbolic periodic orbits at least equal to the largest exponent of the action in homology.

Mathematics Subject Classification (2010). Primary 37-02, 37Axx, 37Cxx, 37Dxx.

Keywords. Geometric structure, ergodicity, partial hyperbolicity, entropy, Lyapunov exponents.

1. Introduction

In this survey we shall present some relations between measure theory and geometric topology in dynamics. This interaction in particular would imply ergodic properties for partially hyperbolic systems and some rigidity properties of actions by higher rank groups. Also we will try to give some problems and directions. We shall not try to be exhaustive on the results, we want just to

*I am indebted with Jana and Raúl for many years of joint contributions. The insight I get of all the subjects discussed here comes in large extent from many discussions and even fights on what is most likely, what is trivial and where is the substance. I hope there are more such discussions in the future, and of course and better, the outcomes coming from there.

IMERL, Facultad de Ingeniería, Universidad de la República, CC 30, Montevideo, Uruguay. E-mail: frhertz@fing.edu.uy.

give some flavor of how this relation works. Moreover, we will concentrate more on possible future directions rather than known results.

The main idea comes as follows, on one hand geometric topology is strong enough so that not any geometric structure can be invariant, but some geometric structure should be preserved by the dynamics at least in some weak sense, this comes in general from homotopic information on the system. On the other hand, measure theory is soft enough that an invariant geometric structure almost always appears along some carefully chosen invariant measure.

One of the principal examples of these invariant geometric structures are invariant stable and unstable manifolds. One may have them appearing in a uniform fashion like in Anosov systems, or in a more nonuniform way. Let us put an example of what we mean:

Take a C^∞ diffeomorphism f of a manifold and assume that its action in homology has spectral radius larger than one. This implies, by Yondim's solution [76] of Shub entropy conjecture [69], that the topological entropy of f is positive, $h_{top}(f) > 0$. By the variational principle there is an ergodic invariant measure μ with positive entropy $h_\mu(f) > 0$. Using Ruelle's inequality [64], $h_\mu(f) \leq \sum_{\lambda_i > 0} \lambda_i$, we get that μ has at least one positive Lyapunov exponent, indeed at least one positive and one negative. Using Pesin theory [5] this implies that there are stable and unstable manifolds associated to negative and positive exponents, they form some kind of absolutely continuous measurable foliations with smooth leaves. Moreover, one may parametrize the unstable manifolds to write the dynamics in some normal form similar to the normal forms around expanding fixed points [29], if no resonance between Lyapunov exponents appears then this normal forms become affine. When the dimension of the manifold is 2 then the measure is hyperbolic, i.e. has no zero Lyapunov exponents and hence, by Katok's shadowing lemma [33], the system is full of hyperbolic orbits and indeed of hyperbolic sets with topological entropy as close to the topological entropy of the diffeomorphism as wanted, in particular the asymptotic growth of hyperbolic periodic orbits is at least equal to entropy and hence at least log of the largest eigenvalue in homology. Getting exact multiplicative estimates in the growth is still an open problem which is also related to the finiteness of entropy maximizing measures.

We conjecture that something similar should be true in higher dimensions. Indeed, in dimension 3 for example, we conjecture that if a smooth diffeomorphism has no hyperbolic periodic orbit (or only finitely many) then the action in first homology has at most one positive and one negative Lyapunov exponent.

When the invariant geometric structures are continuous there is some hope to get more, and even of classifying the systems. The hallmark of simply described system is the affine automorphism of a homogeneous space, i.e. when $M = G/H$ is the quotient of a Lie group G by a closed subgroup H, and $f = gL$, where L is an automorphism of G leaving H invariant and g is any element in G. When the system preserve some continuous foliations, it is expected that the dynamics be conjugated, or at least come in some standard

type of construction from these affine models. But this is far from understood even for the uniformly hyperbolic, Anosov systems. On the other hand, for expanding maps the picture was completely understood since the early 80's by the results of M. Shub [68] and M. Gromov [25] and expanding systems are always conjugated to expanding automorphisms on nilmanifolds.

There are lot of unsolved issues even for perturbations of these affine models. One of these being stable ergodicity and robust transitivity for perturbations of affine automorphisms,local rigidity and measure rigidity for higher rank actions. Growth of the number of compact invariant submanifolds, even understanding w.r.t. to what should be measured the growth is a nontrivial fact (volume, diameter, homology complexity...). These issues will hopefully appear along the rest of the paper.

2. Rigidity of Actions of Higher Rank Groups

Let us begin with the simplest model, multiplication by an integer on the circle. We shall see how to get global rigidity for commuting covering maps of degree larger than one on the circle, see [51] for details. In this case, degree is used to get a semi-conjugacy with the linear map, and hence preservation of some structure, then this semi-conjugacy is used to get a measure which is large, i.e. projects to Lebesgue measure under the semi-conjugacy and hence with positive entropy. The measure is used to find a periodic point with some non-collapsing property. This implies that Lyapunov exponents for this periodic orbits are multiple of the linear exponents.

Let us state an interesting corollary of the general theorem. Let $f, g : S^1 \to S^1$ be two C^1 maps of degree 2 and 3 respectively, i.e. $\deg(f) = 2$ and $\deg(g) = 3$ Assume also that f has at most finitely many critical points (this includes having no critical point at all).

Theorem 2.1. *Let f and g be as above, then there is a C^1 diffeomorphism $h : S^1 \to S^1$ such that $h \circ f = 2h$ and $h \circ g = 3h$, i.e. f and g are smoothly conjugated to the linear action.*

The theorem says in particular that there are no critical points neither for f nor for g and moreover f and g should be expanding maps with an absolutely continuous invariant measure. We believe that the assumption on finitely many critical points can be removed. Previous results in this direction used that the maps were expanding and required at least $C^{1+\alpha}$ smoothness in order to get existence and uniqueness of smooth absolutely continuous invariant measures. In [78] appeared a proof for expanding maps without the use of absolutely continuous invariant measure for expanding maps close to ours. In the general C^r case, $1 < r \leq \infty, \omega$ there is also conjugacy of the same smoothness as the action. This already follows from a result by A. Johnson and D. Rudolph, [28]. Previous result by Sacksteder [65] obtained the conjugacy but

with some loss of smoothness. M. Shub and D. Sullivan [71] proved that two expanding maps on the circle with same homotopy data are smoothly conjugated without any loss of derivatives. For C^1 expanding maps there need not be an absolutely continuous measure [37], and even if it exists one may not have uniqueness [50]. With our method and using Sternberg local linearization theorem [73] we can recover all the other proofs in addition to get the C^1 general case.

Let us emphasize here the feature that the theorem holds for C^1 actions though we make use of some sort of Pesin theory. As is well established, some Hölder condition on the derivative is needed in order to use Pesin theory, see [47]. Most rigidity results makes crucial use of Pesin theory. Here with C^1 is enough, what makes it work in the circle is that the unstable manifold of a point is already given, it is an interval. In higher dimension the unstable manifold is not given a priori in general, what makes the use of all the power of Pesin theory unavoidable at first glance. One may wonder what will happens if unstable manifolds are already given with some regularity, like in geodesic flows, where horospheres comes from geometry. Even in the case one gets a foliation by horospheres, its regularity may be quite bad, for example, in nonpositive curvature it is not known if it is absolutely continuous. It is know that it is not Hölder continuous for large Hölder exponent at some points, [23], but maybe with larger systems one may improve regularity. In any event, since the complete picture is not known in the general setting, there is no harm on assuming any smoothness to begin with.

It is more striking the fact that no dynamical assumption is given in the circle case, i.e. no expanding assumption. But by now we assumed finitely many critical points and though this does not look like a dynamical assumption, it turns out to be dynamical. It is this assumption which gives as unstable manifolds for the orbits, if not we will need to use a $C^{1+\alpha}$ hypothesis and will get also rigidity.

In higher dimensions, there are some results on this direction, but there are examples of actions with Cartan homotopy data but not conjugated to the linear ones. Here by homotopy data we mean the action in the first homotopy group and by Cartan we mean maximal rank with all elements hyperbolic. Let us observe that after the global rigidity results in [53] we get that the non-conjugated actions homotopic to linear Cartan actions on tori can not be C^1 close to the linear ones and cannot even be simultaneously homotoped to the linear action. We believe that on one hand this is a problem at a C^0 level, i.e. if the semiconjugacy is injective then it is a diffeomorphism, on the other hand we also believe that only finitely many blowups can be carried out and hence not much harm can be done.

In [36] we proved, with A. Katok, that a real analytic action of $SL(n,\mathbb{Z})$ on $\mathbb{T}^n$ preserving some non trivial measure and with standard homotopy data is essentially conjugated to the linear one. More precisely,

Theorem 2.2. *Let $\Gamma \subset SL(n,\mathbb{Z})$, $n \geq 3$ be a finite index subgroup. Let ρ be a C^ω (real-analytic) action of Γ on $\mathbb{T}^n$ with standard homotopy data, preserving an ergodic measure μ whose support is not contained in a ball. Then:*

1. *There is a finite index subgroup $\Gamma' \subset \Gamma$, a finite ρ_0-invariant set F and a bijective real-analytic map*

$$H : \mathbb{T}^n \setminus F \to D$$

where D is a dense subset of $\operatorname{supp}\mu$, such that for every $\gamma \in \Gamma'$,

$$H \circ \rho(\gamma) = \rho_0(\gamma) \circ H.$$

2. *The map H^{-1} can be extended to a continuous (not necessarily invertible) map $P : \mathbb{T}^n \to \mathbb{T}^n$ such that $\rho \circ P = \rho_0 \circ P$. Moreover, for any $x \in F$, pre-image $P^{-1}(x)$ is a connected set.*

3. *For $\Gamma = SL(n,\mathbb{R})$ one can take $\Gamma' = \Gamma = SL(n,\mathbb{R})$.*

The main issues in proving theorem 2.2 are existence of semiconjugacy with the linear model, [46], existence of a periodic orbit for the action, or equivalently a fix point for a finite index subgroup and linearization of the action around the fixed point, [18]. For the existence of a periodic orbit we rely on our earlier result about measure rigidity [34] where uniqueness of a large measure is proven for actions with Cartan homotopy data. Previously B. Kalinin and A. Katok [30], proved that such a large measure should be absolutely continuous w.r.t. Lebesgue measure. In [32] we extended the non-uniform measure rigidity result to a general setting only putting assumptions on Lyapunov exponents of the system and entropy.

Opposite from single diffeomorphisms, existence of periodic orbits for actions of larger groups is a completely open area, very little is known, even for flows and commuting diffeomorphisms. It is not clear what kind of invariants should guarantee the existence of periodic orbits, something that generalize Lefschetz formula and indexes of fixed points. Of course one may think on compact leaves of foliations also and their stability.

For example, let ρ be an action of $SL(n,\mathbb{Z})$ on an $n-$dimensional ball. Should it have a periodic orbit? should it have a periodic orbit in its interior? what about if the action in the boundary sphere is the projective one. See the related Thurston's stability result in [74]. What about linearization of smooth actions around fixed points? see [18] for some results along this directions.

Let us finish with a problem on analytic functions:

Problem 1. *Let K_1 and K_2 be two compact subsets of $[0,1]$ with positive measure and let $A = K_1 \times [0,1] \cup [0,1] \times K_2$. Let $\phi : A \to \mathbb{R}$ be a continuous map. For every $x \in K_1$ define $\phi_x : [0,1] \to \mathbb{R}$ to be $\phi_x(y) = \phi(x,y)$. Similarly*

define ϕ^y for $y \in K_2$. Assume that ϕ_x is real analytic with radius of convergence 2 for every $x \in K_1$ and $\phi_y : [0,1] \to \mathbb{R}$ is real analytic with radius of convergence 2 for every $y \in K_2$. Prove that ϕ coincides with a real analytic map in some open set U such that $Leb(A \cap U) > 0$.

If necessary, may be assumed that for some density point p of A, the Taylor series of ϕ has positive radius of convergence.

3. Partially Hyperbolic Systems

I would said that one of the most resisting problems in partially hyperbolic systems is the robust transitivity of Kolmogorov affine diffeomorphisms on homogeneous spaces. Indeed there is no natural example of a robustly transitive diffeomorphism. The examples of M. Shub [70], R. Mañé [43], C. Bonatti and L. Díaz [10], are built by a careful perturbation of a simple system. In [10] a mechanism for mixing called blender was created. But though these blenders seem to be abundant in the absence of hyperbolicity, they are not present in most natural examples. It would be interesting to have an example of a robustly transitive time one map of Anosov flow, or even a robustly transitive partially hyperbolic diffeomorphism with isometric central direction. All this examples, when transitive, can be approximated by robustly transitive diffeomorphisms as done in [10] with blenders, but this is by now the only known mechanism.

A very important issue in the classification problem for partially hyperbolic systems is the integrability of the center distribution. It is known that the center distribution is not always integrable, see the paper by K. Burns and A. Wilkinson [15] for a nice account. The lack of integrability comes from two sources, one is the break of Froebenius condition on integrability of a bundle and the other is the lack of differentiability of the bundle. For a while the only known type of example were the ones coming from the break of Froebenius condition. Indeed, for some time we believed that when the dimension of the center bundle is one (and hence any notion of Froebenius condition should hold) then it should integrate to a foliation. But recently we found the example in [59], whit the center bundle having no foliation tangent to it. The example is an Axiom A partially hyperbolic diffeomorphism on $\mathbb{T}^3$ and indeed is a quite simple one. Of course it would be much more interesting to have a transitive example and we hope there is not such example. It is worth comparing this example with the results of M. Brin, D. Burago and S. Ivanov, [12, 13], where it is proven in particular that on $\mathbb{T}^3$ the center bundle is always uniquely integrable if there is a stronger partially hyperbolic condition.

One of the driving directions in the theory of partially hyperbolic systems has been the search of ergodicity and criteria guaranteeing it. Indeed, Pugh and Shub conjectured, [49]

Conjecture 1. *Stable ergodicity is open and dense among smooth volume preserving partially hyperbolic systems.*

They gave a program for proving the conjecture which splits the conjecture in proving that systems having the accessibility property, i.e. any two point can be joined by a path made of stable and unstable leaves (an su-path), are on one hand ergodic and on the other hand such systems contain an open and dense set. For the first part, K. Burns and A. Wilkinson, [14] completely solved the problem under some assumption called center-bunching which essentially means that the dynamics in the central is close to conformal when compared with the contraction and expansion rates of the strong directions. In particular, when the central dimension is 1 this center bunching condition is automatically satisfied. In this case we gave another proof in [55] where we also give a solution to the Pugh-Shub stable ergodicity conjecture when the central dimension is 1.

In [60], with J. Rodriguez Hertz, A. Tahzibi and R. Ures, new criteria for ergodicity w.r.t. volume measure of smooth systems is given (partially hyperbolic or not). Indeed, it is given a description of the ergodic components in Pesin theory, this are parallel to the homoclinic classes of hyperbolic periodic points. We attach to each hyperbolic periodic point p (and in fact to its homoclinic class) a set called Pesin homoclinic class which consists of regular points which has some nontrivial transverse intersection of stable or unstable manifolds to the ones of p. To be more precise, define

$$B^s(p) = \{x \text{ s.t. } W^s(x) \pitchfork W^u(o(p)) \neq \emptyset\},$$

here p is assumed to be a forward regular point, similarly is defined $B^u(p)$ with backward regular points. Our criteria says

Theorem 3.1. *If f is volume preserving and $B^s(p)$ and $B^u(p)$ are of positive volume then they coincide a.e. with its intersection $B(p)$ and the dynamics on this set is ergodic and non-uniformly hyperbolic.*

Using this criterium and blenders we solved the Pugh-Shub stable ergodicity conjecture when the central dimension is 2 in [60]. It seems likely that this method should be useful to solve Pugh-Shub stable ergodicity conjecture in C^1 topology without any restriction on the central direction.

We think that to prove the conjecture in higher differentiability, a thorough understanding of the structure of the accessibility classes is in order. Indeed this was the way we solved the conjecture with J. Rodriguez Hertz and R. Ures when the central dimension is one [55] and for perturbations of toral automorphisms [52]. When the central direction is one dimensional, the partition into accessibility classes generates a lamination, i.e. the complement of the open accessibility classes is laminated by the accessibility classes. So, to get accessibility one should prove that this lamination cannot exist, or destroy it by

perturbations. In dimension 3, for example, we were able to prove that in some manifolds, this lamination cannot exist and hence every partially hyperbolic system is ergodic. We conjectured [56] the following:

Conjecture 2. *Let M be a manifold which is not the mapping torus of a toral automorphism commuting with an Anosov map. Then any smooth partially hyperbolic diffeomorphism is ergodic.*

The following result in [58] supports the conjecture:

Theorem 3.2. *Let $f : M \to M$ be a homeomorphism of a three dimensional irreducible (every sphere bounds a ball) manifold. Assume there is an embedded torus $T \subset M$ such that $f(T) \sim T$ and $f_{\#}|\pi_1(T)$ is hyperbolic. Then $M \setminus T = \mathbb{T}^2 \times (0, 1)$.*

The theorem implies that the lamination generated by the accessibility classes has no compact leaves unless the manifold is one of the mentioned in conjecture 2. This in turns seems to imply that the lamination should be a minimal foliation of the whole manifolds. To prove this step, one way is to solve the following:

Problem 2. *Let $f : S \to S$ be a diffeomorphism of a complete surface S. Assume f is Anosov, with uniform local product structure and with dense set of periodic orbits. Is S a torus?*

Ones we get that the partition into accessibility classes is a minimal foliation, smoothness of this foliation implies it is ergodic w.r.t. Lebesgue measure, see [20]. But it is still not clear for us how to get the desired smoothness (at least C^2).

On the other hand, the criterium was used by A. Avila and J. Bochi [1] to strengthen R. Mañé, J. Bochi, M. Viana, theorem, [45, 8, 9] to get that for a generic volume preserving diffeomorphism, either there is at least one zero Lyapunov exponent a.e. or the Oseledets splitting extends to a dominated splitting on the whole manifold. Previous results gave that Oseledets splitting extend to a dominated one over closed invariant sets of arbitrary large measure. In dimension 2, this gives that for a generic volume preserving diffeomorphism, either it is Anosov, or all exponents vanish a.e.

J. Rodriguez Hertz, in [54] proved the full conjecture in dimension 3, i.e.

Theorem 3.3. *For volume preserving diffeomorphisms in dimension 3 belonging to a C^1 generic set, either all exponents vanish a.e., or the system is ergodic and non-uniformly hyperbolic with the Oseledets splitting extending to a dominated splitting.*

For dissipative systems, in [63] we extend the criterium for SRB measures and prove uniqueness of SRB measures for transitive surface diffeomorphisms. Existence of SRB measures is still a wide open problem, there are lot of advances, but mostly for particular examples like Henon-like maps. It is in general

believed that having no zero exponents on a set of positive Lebesgue measures should imply existence of an SRB measure, but the examples in [27] are saying that maybe something more is needed to guarantee them at least if we want to work with finite measures.

In our survey [57] on partially hyperbolic systems the reader may find a large list of problems and directions. Even if some new results appear since then we think that it is still updated.

4. Product Measures and Entropy Formula

One of the most striking formulas in ergodic theory is Ledrappier-Young entropy formula, stating that entropy is a linear combination of the positive Lyapunov exponents, where the coefficients have some dimension like meaning. To be more precise, Let f be a $C^{1+\alpha}$ diffeomorphism and μ be an invariant measure. Oseledets theorem gives a splitting

$$TM = E^u_1 \oplus E_2 \oplus \cdots \oplus E^u_u \oplus E_0 \oplus E^s_s \oplus \cdots \oplus E^s_2 \oplus E^s_1$$

associated to the positive, zero and negative Lyapunov exponents

$$\chi^u_1 > \chi^u_2 > \cdots > \chi^u_u > 0 > \chi^s_s > \cdots > \chi^s_2 > \chi^s_1.$$

Call $E^u = E^u_1 \oplus E_2 \oplus \cdots \oplus E^u_u$ and $E^s = E^s_s \oplus \cdots \oplus E^s_2 \oplus E^s_1$, E_0 is sometimes also denoted E^c.

Pesin stable manifold theorem states that tangent to the flag

$$E^u_1 \subset E^u_1 \oplus E^u_2 \subset \cdots \subset E^u_1 \oplus E_2 \oplus \cdots \oplus E^u_u = E^u$$

there is a flag $V_1(x) \subset V_2(x) \subset \cdots \subset V_u(x) = W^u(x)$. In general, the slow bundles E^u_i, $i > 1$ are not integrable, though one may put some locally invariant families of disks tangent to them which are commonly called $W_i(x)$, but are in general not canonically defined. Indeed one may build examples of affine systems where this bundles cannot be integrated, but there are also some cases where the E_i integrates in some canonical way, for example, for toral automorphisms. We call fast invariant manifolds to the V_i's and slow manifolds to the W_i's.

Associated to this flag, there are measurable partitions $\xi_1 > \xi_2 > \cdots > \xi_u$ subordinated to the $V_i's$. So we have conditional measures μ^{ξ_i} for $i = 1, \ldots, u$. Let us call d_i the Haussdorff dimension of the μ^{ξ_i} and $\gamma_i = d_i - d_{i-1}$ $i = 1, \ldots, u$ (here we put $d_0 = 0$. So the γ_i's are some sort of transverse dimensions associated to some transverse measures. It can be seen that the d_i does not depend on the ξ_i but only on the conditional measures along the V_i's, μ^i, and the same for $h_\mu(f, \xi_i)$ which will then be called $h_\mu(f, V_i)$. Ledrappier-Young [40, 41] formula then states that

Theorem 4.1. *Let f and μ be as above, then for every $r = 1, \dots, u$*

$$h_\mu(f, V_r) = \sum_{i=1}^{r} \gamma_i \chi_i^u$$

and for $r = u$,

$$h_\mu(f) = \sum_{\chi_i > 0} \gamma_i \chi_i$$

Shub-Wilkinson [72] gave examples of partially hyperbolic systems with circle center bundle having positive central exponents on $\mathbb{T}^3$. In this case it follows that $\chi_1 > \chi_2 > 0$, $d_1 = 1$, $d_2 = 2$ and $\gamma_1 = \gamma_2 = 1$, the conditional measures along the central foliation are atomic so this conditional dimensions are 0. This is very important to understand that the conditional dimension can be completely different to the transverse direction, though it can be seen that the conditional dimension is alway smaller than or equal to the transverse dimension. In fact the matching of conditional measures and transverse measures give rise to all sorts of rigidity. When the measure μ is Lebesgue measure (or at least some SRB measure), absolute continuity of the slow directions is known to imply, in some cases, that conditionals and transverse dimensions match and in this case some rigidity appears. See [4], [24] and [75] for examples of this features. In particular, in [4] it is proven the rigidity property conjectured in our survey [57]. To get this rigidity it comes into play the pioneering work of F. Ledrappier [38] on random product of matrices. It was already there that this type of rigidity appeared. We think that indeed this features should be quite general and rigidity comes from the fact that conditional and transverse dimensions match. Let us formulate the following

Conjecture 3. *If all transverse dimensions coincide with the conditional dimensions, then the measure $\mu^u = \mu_{W^u}$ is locally a product.*

It is a part of the conjecture to make some meaning of the conditional dimension and so the meaning of been locally a product, one should use the W_i's and try to see that this dimensions do not depend on the choices, or make some fix choice. We may assume to begin with some integrable case where the slow directions W_i's become integrable. It would be also a nice exercise to understand first how things work for invariant measures of toral automorphisms.

When all conditional measures (also the ones in slow directions) are Lebesgue, absolute continuity of the fast foliation implies that μ_{W^u} is a product measure and hence it is Lebesgue. F. Ledrappier and J-S. Xue analyzed also the other extreme in [39], they prove that if the transverse dimension γ_i is 0 then the measure μ_i is supported in V_{i-1} and hence equals μ_{i-1}. This problem appeared in [35] when relations between smoothness of entropy map and rigidity of measures was proved.

It is very important to observe that the conditional dimensions does not depend on the system f, just on the measures and the invariant manifolds (fast

or slow), but the transverse dimensions depend on how this invariant manifolds fit inside the flag. For example, if two systems leave invariant the same measure μ and have the same Oseledets splitting, then a priori the $\gamma_i's$ may change since the flag may change. It is non trivial to see that this does not happens in some cases and this was done by H. Hu in [26] where essentially it is proven that the γ_i's does not change as long us they correspond to expanding Lyapunov exponents for both systems. In [26] the systems are assumed to commute but we believe that this is almost essentially the case if the Oseledets splitting coincides.

Another related problem is to understand when the measure μ itself is locally a product of μ^u, μ^c and μ^s. When the three conditional measures are absolutely continuous w.r.t. Lebesgue and the central direction is associated to some absolutely continuous foliation the measure itself is absolutely continuous w.r.t. to Lebesgue, [41]. In the general case, almost nothing is known. The key issue here seems to be absolute continuity of the complementary foliation w.r.t. the conditional measures. But there is not much advances in this directions, not even for linear maps. This again is a source of rigidity. Let us point out that here we are looking for measures that are product (or equivalent to product measures) and not only that they possesses asymptotically almost local product structure, this was already done in the work of L. Barreira, Y. Pesin and J. Schmeling, [6] where it is proven that every hyperbolic measure is assymptotically a product.

Observe also that in the Furstenberg measure rigidity problem for liner automorphisms on $\mathbb{T}^3$, the issue is with a measure having 0 conditional dimensions on all three directions, but is not known if it is locally a product. When some conditional dimension is non-zero then the problem is completely solved. If the group is large enough then, it can be seen that the system has already non-zero conditional dimension. This was already done in [11] and [7] for the case of non abelian groups. In the case of abelian semigroups acting on the circle in [21] M. Einsiedler and A. Fish assume that the semigroup growth with some positive rate. Let us put the following problem which may be seen as a step in the middle of Furstenberg problem and [21].

Problem 3. *Let Γ be a non finitely generated multiplicative subsemigroup of $\mathbb{N}$, then the unique ergodic invariant measures for the linear action are either Lebesgue or atomic.*

In a conversation with V. Bergelson, it appeared that the first example not included in previous works that may have a chance is a semigroup generated by some increasing powers of the prime numbers.

Other instance where a product structure of the measure appears is for entropy maximizing measures. In [62] we prove that

Theorem 4.2. *For partially hyperbolic systems in dimension 3 with compact central foliation having the accessibility property, either there is only one entropy maximizing measure and the system is conjugated to an isometric ex-*

tension of an Anosov system, or there are at least two entropy maximizing measures, one with negative central exponents and the other with positive ones.

This theorem relies on a theorem of A. Avila and M. Viana [3] for cocycles over hyperbolic dynamics with invariant product measures, extending the mentioned F. Ledrappier [38] result, and is parallel to the results in [4] where similar results are proven for Lebesgue measure and relies in the work of [2]. We conjecture that if the system is transitive then there are at most 2 entropy maximizing measures, this should follow from some averaged central dynamics. For the examples of partially hyperbolic systems in dimension 3 isotopic to Anosov diffeomorphisms it is proven in [17] that there is only one measure of maximal entropy. It would be interesting to know what happens for perturbations of the time one map of the geodesic flow over hyperbolic surfaces, we believe that something similar to our theorem 4.2 happens there. We think that the following should also be true, this conjecture is close to the one appearing in the work of J. Buzzi [16]

Conjecture 4. *If f is a C^∞ transitive surface diffeomorphism with positive topological entropy then there is only one measure of maximal entropy. Moreover, this measure is locally a product.*

That the measure is locally a product seems likely to follow from the method of construction appearing in [16], i.e. building countable Markov chains associated to the dynamics, what is not clear is how to build this countable Markov chain. So it is reasonable to conjecture that entropy maximizing measures are always locally a product in dimension 2.

In [3] A. Avila and M. Viana combine the technique of cocycle rigidity with our previous result [52] on stable ergodicity for automorphisms on tori with 2 dimensional center bundle to get stably Bernoulli for symplectic perturbations. Indeed they prove a dichotomy along ours, getting that either the system is nonuniformly hyperbolic in the case the system has the accessibility property or the system is smoothly conjugated to the linear one. We think that there may be a proof of stable ergodicity for linear automorphisms regardless of its central dimension combining these ideas. To be a little bit more precise, it seems that it can be proved that every accessibility class is dense. In case f has not the accessibility property the problem is that the s- and u-bundles need not be jointly integrable, still we think that they fit inside some integrable subbundle and this should coincide with the bundle of nonzero exponents.

On the other hand one can see this result as a nonlinear version of the result in [42]. There it is proven in particular that any ergodic measure invariant by an irreducible linear automorphism of the torus is either Lebesgue measure or the support of the central conditional measures is a circle. One can also apply this to study compact invariant subsets. The same technic applies to some extent to study partially hyperbolic systems with isometric central dynamics. We believe that with some zero Lyapunov exponents assumption on the central direction, or even some zero entropy assumption on the central direction (whatever this

means) some close results should be expected. In any case, the simplest case of the time one map of Anosov flows seems to be not clear. Indeed the following problem by J. Rodriguez Hertz and F. Ledrappier is still open.

Problem 4. *Are there nonfinite compact minimal invariant sets invariant by the time one map of an Anosov flow but not for the flow?*

They posed the question for the case of the geodesic flow of a surface of constant curvature. It seems likely that the answer is yes for most systems, indeed, we think that an example for the geodesic flow of some constant curvature surface is also doable playing with the length of geodesic of a pair of pants, or equivalently using the Teichmüller flow, but it is not clear for every surface. In any case, how do these sets look like, can they be geodesic laminations? what about entropy of this sets, even if the set is not minimal? The same question for invariant measures seems to be true in all cases (for Anosov flows).

References

[1] A. Avila, J. Bochi, *N*onuniform Hyperbolicity, Global Dominated Splittings and Generic Properties of Volume-Preserving Diffeomorphisms. (preprint)

[2] A. Avila, J. Santamaria, and M. Viana. *C*ocycles over partially hyperbolic maps. (preprint)

[3] A. Avila and M. Viana. *E*xtremal Lyapunov exponents: an invariance principle and applications. preprint

[4] A. Avila, M. Viana, and A. Wilkinson, *A*bsolute continuity, Lyapunov exponents, and rigidity. In preparation.

[5] L. Barreira and Ya. Pesin, *Nonuniform Hyperbolicity: Dynamics of Systems with Nonzero Lyapunov Exponents*, Encyclopedia of Mathematics and Its Applications, **115** Cambridge University Press, 2007.

[6] L. Barreira; Ya. Pesin and J. Schmeling, *Dimension and product structure of hyperbolic measures.* Ann. of Math. (2) 149 (1999), no. 3, 755–783.

[7] Y. Benoist and J-F Quint, *M*easures stationnaires et ferme's invariants des espaces homnogenes (preprint)

[8] J. Bochi, *Genericity of zero Lyapunov exponents.* Ergodic Theory Dynam. Systems 22 (2002), 1667–1696.

[9] J. Bochi, M. Viana, *The Lyapunov exponents of generic volume-preserving and symplectic maps.* Ann. of Math. (2) 161 (2005), no. 3, 1423–1485.

[10] C. Bonatti, L. Diaz, *P*ersistent nonhyperbolic transitive diffeomorphisms. Ann. of Math. (2) 143 (1996), no. 2, 357–396

[11] J. Bourgain, A. Furman, E. Lindenstrauss, S. Mozes, *I*nvariant measures and stiffness for non-abelian groups of toral automorphisms. C. R. Math. Acad. Sci. Paris 344 (2007), 737–742.

[12] M. Brin, D. Burago, S. Ivanov *Dynamical coherence of partially hyperbolic diffeomorphisms of the 3-torus.* J. Mod. Dyn. 3 (2009), no. 1, 1–11.

[13] D. Burago, S. Ivanov *Partially hyperbolic diffeomorphisms of 3-manifolds with abelian fundamental groups.* J. Mod. Dyn., Vol. 2, 541–580, 2008.

[14] K. Burns, A. Wilkinson, *On the ergodicity of partially hyperbolic diffeomorphisms.* to appear in Annals of Math.

[15] K. Burns, A. Wilkinson, *Dynamical coherence and center bunching.* Discrete and Continuous Dynamical Systems, (Pesin birthday issue) 22 (2008), 89–100.

[16] J. Buzzi, *M*aximal entropy measures for piecewise affine surface homeomorphisms. Ergodic Theory and Dynamical Systems (2009), **29**, 1723–1763.

[17] J. Buzzi, T. Fisher, M. Sambarino, C. Vazquez, *M*aximal Entropy Measures for certain Partially Hyperbolic, Derived from Anosov systems. preprint

[18] G. Cairns, E. Ghys, *The local linearization problem for smooth $SL(n)$-actions.* Enseign. Math., II. Ser. 43, No.1–2, 133–171 (1997).

[19] R. de la Llave, *Smooth conjugacy and SRB measures for uniformly and non-uniformly hyperbolic systems.*, Comm. Math. Phys. **150** (1992), 289–320.

[20] B. Deroin, V. Kleptsyn, A. Navas, *On the question of ergodicity for minimal group actions on the circle.* preprint arXiv:0806.1974

[21] M. Einsiedler and A. Fish, *Rigidity of measures invariant under the action of a multiplicative semigroup of polynomial growth on* $\mathbb{T}$. Ergodic Theory Dynam. Systems.

[22] J. Franks, *Invariant sets of hyperbolic toral automorphisms*, Amer. J. Math. (1977), 99. 1089–1095.

[23] M. Gerber, V. Niţică, *H*ölder exponents of horocycle foliations on surfaces. Ergodic Theory and Dynamical Systems (1999), **19**, 1247–1254.

[24] A. Gogolev *H*ow typical are pathological foliations in partially hyperbolic dynamics: an example. preprint

[25] M. Gromov, *Groups of polynomial growth and expanding maps.* Inst. des Hautes Études Sci. Publ. Math. **53**, (1981), 53–73.

[26] H. Hu *Some ergodic properties of commuting diffeomorphisms*, Ergodic Theory Dynam. Systems **13** (1993), no. 1, 73–100.

[27] H. Hu and L.-S. Young, *Nonexistence of SRB measures for some systems that are almost Anosov.* Ergodic Theory and Dynamical Systems 15 (1995), 67–76.

[28] A. Johnson, A, D. Rudolph, *C*ommuting endomorphisms of the circle. Erg. Th. and Dynam. Sys. **12** (1992), no. 4, 743–748.

[29] B. Kalinin and A. Katok, *Invariant measures for actions of higher rank abelian groups*, Proc. Symp. Pure Math, **69**, (2001), 593-637.

[30] B. Kalinin and A. Katok, *Measure rigidity beyond uniform hyperbolicity: Invariant Measures for Cartan actions on Tori*, Journal of Modern Dynamics, **1** N1 (2007), 123–146.

[31] B. Kalinin, A. Katok and F. Rodriguez Hertz, *New Progress in Nonuniform Measure and Cocycle Rigidity* Electronic Research Announcements in Mathematical Sciences, **15**, (2008), 79–92.

[32] B. Kalinin, A. Katok and F. Rodriguez Hertz, *Nonuniform Measure Rigidity*, preprint (2008), arXiv:0803.3094

[33] A. Katok, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms.* Publ. Math. IHES No. **51**, (1980), 137–173.

[34] A. Katok and F. Rodriguez Hertz, *Uniqueness of large invariant measures for $\mathbb{Z}^k$ actions with Cartan homotopy data*, Journal of Modern Dynamics, **1**, N2, (2007), 287–300.

[35] A. Katok and F. Rodriguez Hertz, *M*easure and cocycle rigidity for certain non- uniformly hyperbolic actions of higher rank abelian groups. preprint, arXiv:1001.2473

[36] A. Katok and F. Rodriguez Hertz, *Rigidity of real-analytic actions of $SL(n,\mathbb{Z})$ on $\mathbb{T}^n$: a case of realization of Zimmer program.* Discrete and Continuous Dynamical Systems **27**, N2, (2010).

[37] K. Krzyżewski *A* remark on expanding mappings. Colloq. math **XLI** (1979), 291–295.

[38] F. Ledrappier *P*ositivity of the exponent for stationary sequences of matrices. In Lyapunov exponents (Bremen, 1984), v. 1186 of Lect. Notes Math., pages 56–73. Springer, 1986.

[39] F. Ledrappier, J-S, Xue, *On the extremal values of transverse dimension in Pesin theory*

[40] F. Ledrappier, L.-S. Young, *The metric entropy of diffeomorphisms. I. Characterization of measures satisfying Pesin's entropy formula*, Ann. Math. **122**, no. 3, 509–539.

[41] F. Ledrappier, L.-S. Young. *The metric entropy of diffeomorphisms. Part II: relations between entropy, exponents and dimension*, Ann. Math. **122** (1985), no. 3, 540–574.

[42] E. Lindenstrauss, K. Schmidt, *Invariant sets and measures of nonexpansive group automorphisms.* Israel J. Math., **144**, 29–60, 2004.

[43] R. Mañé, *Contributions to the C^1-stability conjecture.* Topology **17** (1978), 386–396.

[44] R. Mañé, *Orbits of paths under toral automorphisms.* Proc. Amer. Soc. (1979), 73, 121–125.

[45] R. Mañé, *The Lyapunov exponents of generic area-preserving diffeomorphisms.* International Conference on Dynamical Systems (Montevideo, 1995), Pitman Res. Notes Math. Ser. **362** (1996), 110–119, Longman, Harlow.

[46] G. Margulis, N. Qian, *Rigidity of weakly hyperbolic actions of higher real rank semisimple Lie groups and their lattices.* Ergodic Theory Dynam. Systems **21** (2001), no. 1, 121–164.

[47] C. Pugh; *The $C^{1+\alpha}$ hypothesis in Pesin theory.* Inst. Hautes Études Sci. Publ. Math. **59** (1984), 143–161.

[48] C. Pugh; M. Shub, *Ergodic Attractors.* Transactions of the American Mathematical Society, Vol. **312**, No. 1. (Mar., 1989), pp. 1–54.

[49] C. Pugh; M. Shub, *Stable ergodicity and partial hyperbolicity.* International Conference on Dynamical Systems (Montevideo, 1995), Pitman Res. Notes Math. Ser. **362**, (1996) 182–187, Longman, Harlow.

[50] A. Quas, *Non-ergodicity for C^1 expanding maps and g-measures.* Ergodic Theory Dynamical Systems **16**, (1996) 531–543.

[51] F. Rodriguez Hertz, *Global Rigidity for abelian semigroups actions on the circle.*

[52] F. Rodriguez Hertz, *Stable ergodicity of certain automorphisms of the torus.* Ann. of Math. **162** (1) (2005), 65–107.

[53] F. Rodriguez Hertz, *Global rigidity of certain abelian actions by toral automorphisms.* Journal of Modern Dynamics, **1**, N3, (2007), 425–442.

[54] M. Rodriguez Hertz, *Continuity of the Oseledets splitting for generic conservative diffeomorphisms of 3-manifolds.*

[55] F. Rodriguez Hertz, M. Rodriguez Hertz, R. Ures, *Accessibility and stable ergodicity for partially hyperbolic diffeomorphisms with 1d-center bundle.* Invent. Math. **172**, 2 (2008) 353–381.

[56] F. Rodriguez Hertz, M. Rodriguez Hertz, R. Ures, *Partial hyperbolicity and ergodicity in dimension three.* J. Mod. Dyn. 2 (2008), no. 2, 187–208.

[57] F. Rodriguez Hertz, M. Rodriguez Hertz, R. Ures, *A survey of partially hyperbolic dynamics.* Partially hyperbolic dynamics, laminations, and Teichmüller flow, 35–87, Fields Inst. Commun., **51**, Amer. Math. Soc., Providence, RI, 2007.

[58] F. Rodriguez Hertz, M. Rodriguez Hertz, R. Ures, *Invariant tori with Anosov dynamics in dimension* 3.

[59] F. Rodriguez Hertz, M. Rodriguez Hertz, R. Ures, *A non-dynamically coherent example in $\mathbb{T}^3$.*

[60] F. Rodriguez Hertz, M. Rodriguez Hertz, A. Tahzibi, R. Ures, *New critertia for ergodicity and non-uniform hyperbolicity.* preprint, arXiv:0907.4539

[61] F. Rodriguez Hertz, M. Rodriguez Hertz, A. Tahzibi, R. Ures, *A criterion for ergodicity of non-uniformly hyperbolic diffeomorphisms. ERA-MS* **14**, (2007) 74–81.

[62] F. Rodriguez Hertz, M. Rodriguez Hertz, A. Tahzibi, R. Ures, *On the finiteness of entropy maximizing measures for some partially hyperbolic systems.*

[63] F. Rodriguez Hertz, M. Rodriguez Hertz, A. Tahzibi, R. Ures, *Uniqueness of SRB measures for transitive diffeomorphisms of surfaces.*

[64] D. Ruelle *E*rgodic theory of differentiable dynamical systems. Inst. Hautes Etudes Sci. Publ. Math. 50 (1979), 27–58.

[65] R. Sacksteder, *A*belian semi-groups of expanding maps. Springer Lecture Notes Math., **318**, (1972), 235–248.

[66] R. Saghin, Z. Xia, *Geometric expansion, Lyapunov exponents and foliations.* Annales de l'Institut Henri Poincare (C) Non Linear Analysis **26**, 2, (2009), 689–704.

[67] N. Shah, *Limiting distributions of curves under geodesic flow on hyperbolic manifold.* 25 pages. arXiv:0708.4093v1

[68] M. Shub, *Endomorphisms of Compact Differentiable Manifolds.* Amer. J. Math. **XCI** (1969), 175–199.

[69] M. Shub, *D*ynamical Systems, Filtrations and Entropy. Bull. Amer. Math. Soc. **80** (1974), 27–41.

[70] M. Shub, *T*opologically transitive diffeomorphisms on $\mathbb{T}^4$. Lect. Notes on Math. **206** (1971), 39.

[71] M. Shub and D. Sullivan, *E*xpanding endomorphisms of the circle revisited. Ergod. Th. Dynam. Sys., **5**, (1985), 285–289.

[72] M. Shub, A. Wilkinson, *P*athological foliations and removable zero exponents. Invent. Math. **139** (3) (2000), 495–508.

[73] S. Sternberg, *L*ocal contractions and a theorem of Poincaré. Amer. J. Math. **7**9 (1957) 809–824.

[74] W. Thurston, *A generalization of the Reeb stability theorem.* Topology **13**, (1974), 347–352.

[75] M. Viana, J. Yang, *S*RB measures and absolute continuity for one-dimensional center direction

[76] Y. Yomdin, *Volume growth and entropy.* Israel J. Math. **57** (1987), 285–301.

[77] A. Zeghib. *Sur une notion dautonomie de systèmes dynamiques, appliqu ée aux ensembles invariants des flots dAnosov algébriques.* Ergodic Theory Dynam. Systems, 15(1):175–207, 1995.

[78] M. Zhang, *Smooth linearization and centralizers for expanding maps of the circle.* "Proceedings of the Int. Conf. on Dyn. Sys. and Chaos (Eds., N. Aoki, et al), Vol. 1, (1995), 302–305.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Unique Ergodicity for Infinite Measures

Omri M. Sarig*

Abstract

We survey examples of dynamical systems on non–compact spaces which exhibit measure rigidity on the level of infinite invariant measures in one or more of the following ways: all locally finite ergodic invariant measures can be described; exactly one (up to scaling) admits a generalized law of large numbers; the generic points can be specified. The examples are horocycle flows on hyperbolic surfaces of infinite genus, and certain skew products over irrational rotations and adic transformations. In all cases, the locally finite ergodic invariant measures are Maharam measures.

Mathematics Subject Classification (2010). Primary 37A40, Secondary 37A17

Keywords. Unique ergodicity, Infinite ergodic theory, Horocycle flows, Infinite genus

1. Introduction

1.1. Motivation. A continuous map T on a compact metric space Ω_0 is called uniquely ergodic if it has exactly one invariant probability measure. It is natural to ask what is the right notion of "unique ergodicity" for maps on non–compact spaces whose invariant measures are all infinite. The question is not what is possible, but rather what happens for "natural" examples.

Following a program initiated in [ANSS], we studied the measure rigidity of non–compact analogues of classical uniquely ergodic systems. The systems we studied include horocycle flows on surfaces of infinite genus, and non–compact group extensions of irrational rotations and adic transformations. The purpose of this lecture is to present our findings, and indicate some open problems.

This work grew out of the vision of Jon Aaronson, and it is with great pleasure that I dedidate this paper to him, on the occasion of his birthday.

*Faculty of Mathematics and Computer Science, Weizmann Institute of Science, POB 26, Rehovot, 76100 ISRAEL. E-mail: omsarig@gmail.com.

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802 USA. E-mail: sarig@math.psu.edu.

1.2. Basic Definitions. Let T be a measurable map on a measurable space $(\Omega, \mathscr{B})$, and suppose μ is a σ–finite measure on $(\Omega, \mathscr{B})$ s.t. $\mu(\Omega) \neq 0$. We say that μ is *invariant*, if $\mu(T^{-1}E) = \mu(E)$ for all $E \in \mathscr{B}$. We say that μ is *ergodic*, if for every set $E \in \mathscr{B}$ s.t. $T^{-1}(E) = E$, either $\mu(E) = 0$ or $\mu(\Omega \setminus E) = 0$.

We say that μ is *conservative*, if for every $W \in \mathscr{B}$ s.t. $\{T^{-n}(W)\}_{n\geq 0}$ are pairwise disjoint, $\mu(W) = 0$. This condition is always satisfied in the following cases: (1) μ is a finite invariant measure; and (2) μ is σ–finite non–atomic ergodic invariant measure and T is a bimeasurable bijection [A1].

The ergodic theorems describe the information such measures contain on the almost sure behavior of orbits $\{T^k\omega\}_{k\geq 0}$ ($T^k := T \circ \cdots \circ T$, k times):

Theorem 1.1 (Birkhoff). *Let μ be a finite ergodic invariant measure for $T : \Omega \to \Omega$, then for every $f \in L^1(\Omega, \mathscr{B}, \mu)$, $\frac{1}{N}\sum_{k=1}^{N} f(T^k\omega) \xrightarrow[N\to\infty]{} \frac{1}{\mu(\Omega)}\int_\Omega f d\mu$ μ–a.e.*

Theorem 1.2 (Hopf). *Suppose μ is a σ–finite conservative ergodic invariant measure, then for every $f, g \in L^1(\Omega, \mathscr{B}, \mu)$ s.t. $g \geq 0$ and $\int_\Omega g d\mu > 0$,*

$$\frac{\sum_{k=1}^{N} f(T^k\omega)}{\sum_{k=1}^{N} g(T^k\omega)} \xrightarrow[N\to\infty]{} \frac{\int_\Omega f d\mu}{\int_\Omega g d\mu} \quad \text{for } \mu\text{–almost every } \omega.$$

Specializing to the case when f and g are indicator functions of sets F, G of positive finite measure, we see that if $\mu(\Omega) = \infty$ then the frequency of visits of $T^n(\omega)$ to F and G tends to zero, but the ratio of these frequencies tends to a definite limit.

The limit depends on μ, although it is the same for proportional measures. It is therefore of great interest to know what are the possible ergodic invariant measures up to scaling. To avoid pathologies (cf. [Sch2]), we restrict our attention to measures which are locally finite in some sense which we now make precise.

The following set-up is not the most general possible, but suffices for our purposes. Suppose Ω_0 is a locally compact second countable metric space with Borel σ–algebra $\mathscr{B}_0$. Let $C_c(\Omega_0) := \{f : \Omega_0 \to \mathbb{R} : f \text{ continuous with compact support}\}$. A Borel measure μ on Ω_0 is called a *Radon measure*, if $\mu(C) < \infty$ for every compact set $C \subset \Omega_0$. Equivalently, every $f \in C_c(\Omega_0)$ is absolutely integrable.

In §3.3 we will need to deal with Borel maps T which are only defined on a subset $\Omega \subseteq \Omega_0$, $\Omega \in \mathscr{B}_0$. Let $\mathscr{B} := \{E \cap \Omega : E \in \mathscr{B}_0\}$. A measure μ on $(\Omega, \mathscr{B})$ is called *locally finite*, if $\mu_0(E) := \mu(E \cap \Omega)$ is a Radon measure on Ω_0. If $\Omega = \Omega_0$, then the properties of being Radon and being locally finite are the same.

Theorems 1.1 and 1.2 are almost sure statements. It is interesting to know what are their points of validity.

Definition 1.1. *A point $\omega \in \Omega$ is called* generic *for μ if*

1. *$\mu(\Omega) < \infty$ and for all $f \in C_c(\Omega_0)$, $\frac{1}{N}\sum_{k=1}^{N} f(T^k\omega) \xrightarrow[N\to\infty]{} \frac{1}{\mu(\Omega)}\int_\Omega f d\mu$; or*
2. *$\mu(\Omega) = \infty$, and for all $f, g \in C_c(\Omega_0)$ such that $g \geq 0$ and $\int g d\mu > 0$,*

$$\frac{\sum_{k=1}^{N} f(T^k\omega)}{\sum_{k=1}^{N} g(T^k\omega)} \xrightarrow[N\to\infty]{} \frac{\int_\Omega f d\mu}{\int_\Omega g d\mu}.$$

Our assumptions on Ω_0 and Hopf's theorem guarantee that the set of generic points of a locally finite conservative ergodic invariant measure μ has full μ–measure.

Similar definitions can be made for flows. A *Borel flow* $\varphi : \Omega \to \Omega$ is a group of maps $\varphi^t : \Omega \to \Omega$ ($t \in \mathbb{R}$) such that $(t, \omega) \mapsto \varphi^t(\omega)$ is Borel, and $\varphi^t \circ \varphi^s = \varphi^{t+s}$ ($t, s \in \mathbb{R}$). A Borel measure is called *φ–invariant*, if it is φ^t–invariant for all t. A Borel measure is called *φ–ergodic*, if any Borel set E s.t. $\varphi^{-t}(E) = E$ for all t satisfies $\mu(E) = 0$ or $\mu(\Omega \setminus E) = 0$. A point is called *generic* for a flow, if it satisfies definition 1.1 with $\int_0^N h(\varphi^s\omega)ds$ replacing $\sum_{k=1}^{N} h(T^k\omega)$ ($h = f, g$).

1.3. Measure Rigidity.

Let T be a Borel map on a Borel subset Ω of a second countable locally compact metric space Ω_0. We are interested in the following problems:

1. Find all locally finite T–ergodic invariant measures;
2. Describe their generic points;
3. If there are many measures, find an ergodic theoretic property which singles out just one (up to scaling).

If one or more of these questions can be answered, then we speak (somewhat unorthodoxly) of "measurable rigidity". The strongest form of measure rigidity is unique ergodicity:

Definition 1.2. *T is* uniquely ergodic (u.e.) *if (1) T admits one locally finite invariant measure up to scaling; and (2) every point is generic for this measure.*[1]

It is useful to weaken this as follows. Let δ_y denote the point mass at y. A point ω is called *exceptional* for a map T (resp. a flow φ) if the measure $\sum_{n>0} \delta_{T^n(\omega)}$ (resp. $\int_0^\infty \delta_{\varphi^s(\omega)} ds$) is locally finite.

Definition 1.3. *T is* uniquely ergodic in the broad sense *if (1) up to scaling, T admits one locally finite ergodic invariant measure not supported on a*

[1] Usually unique ergodicity is only defined for continuous maps on compact metric spaces. In this case the unique invariant measure is finite, and (1) implies (2).

single orbit; and (2) every non–exceptional non–periodic point is generic for this measure.

See theorems 2.3 and 2.5 for examples.

Interestingly enough, in the non–compact case there is a large collection of "natural" examples which exhibit a different, more subtle, form of measure rigidity. For these dynamical systems:

- There are no finite invariant measures at all, except perhaps measures supported on periodic orbits;
- There are infinitely many locally finite ergodic invariant measures, all of which can be specified;
- In some cases we know what are the generic points of these measures;
- In some cases we are able show that exactly one of these measures up to scaling admits a generalized law of large numbers (cf. §2.5).

The purpose of this paper is to describe these examples.

2. Horocycle Flows

2.1. Definition. Let M be a complete, connected, orientable hyperbolic surface. Let T^1M be its unit tangent bundle. The *geodesic flow* is the flow $g : T^1M \to T^1M$ which moves a unit tangent vector, at unit speed, along its geodesic. The *Horocycle* of a vector $\omega \in T^1M$ is the set

$$\mathrm{Hor}(\omega) := \{\omega' \in T^1M : \mathrm{dist}(g^s\omega, g^s\omega') \xrightarrow[s\to\infty]{} 0\}. \tag{2.1}$$

We shall soon see that this is a smooth curve in T^1M. The *horocycle flow* of M is the flow $h : T^1M \to T^1M$ which moves $\omega \in T^1M$ at unit speed along $\mathrm{Hor}(\omega)$ in the positive direction.[2]

It is useful to consider the case when $M = \mathbb{H} := \{x + iy : x, y \in \mathbb{R}, y > 0\}$, equipped with the metric $\sqrt{dx^2 + dy^2}/y$. Poincaré's Theorem ([Kat], chapter 1), says that the orientation preserving isometries of $\mathbb{H}$ are Möbius transformations $z \mapsto \frac{az+b}{cz+d}$ where a, b, c, d are real. We denote the collection of these maps by $\mathrm{Möb}(\mathbb{H})$. $\mathrm{Möb}(\mathbb{H})$ acts transitively on $T^1\mathbb{H}$: for every $\omega_1, \omega_2 \in T^1\mathbb{H}$ there exists $\varphi \in \mathrm{Möb}(\mathbb{H})$ s.t. $\varphi_*(\omega_1) = \omega_2$. Schwarz's Lemma says that φ is unique.

Let ω_0 denote the unit tangent vector based at i and pointing north. It is easy to see that the geodesic flow moves ω_0 along the vertical ray it determines. Since every $\omega \in T^1\mathbb{H}$ can be mapped by an element of $\mathrm{Möb}(\mathbb{H})$ to ω_0, and since

[2] Sometimes h is called the *stable* horocycle flow. The unstable horocycle flow is defined in the same way, except that one takes the limit $s \to -\infty$ in (2.1).

isometries map geodesics to geodesics, every geodesic is either a circular arc perpendicular to $\partial\mathbb{H} := \{z : \operatorname{Im} z = 0\}$, or a vertical line.

One can check in a similar way that $\mathrm{Hor}(\omega_0)$ consists of the unit tangent vectors based on the line $\operatorname{Im} z = 1$ and pointing north. Since the hyperbolic metric agrees with the euclidean metric on the line $\operatorname{Im} z = 1$,

$$h^t(\omega_0) = (\psi_t)_*\omega_0, \text{ where } \psi_t : z \mapsto z + t.$$

For general vectors $\omega \in T^1\mathbb{H}$, let φ_ω be the unique element of $\mathrm{Möb}(\mathbb{H})$ s.t. $\omega = (\varphi_\omega)_*\omega_0$, then $\mathrm{Hor}(\omega) = (\varphi_\omega)_*[\mathrm{Hor}(\omega_0)]$ and

$$h^t(\omega) = (\varphi_\omega \circ \psi_t)_*\omega_0. \tag{2.2}$$

The Möbius transformation φ_ω maps the line $\operatorname{Im} z = 1$ onto a circle C which is tangent to $\partial\mathbb{H}$ (possibly at ∞). $\mathrm{Hor}(\omega)$ consists of the unit tangent vectors based at C, perperndicular to C, and pointing in the direction of the tangency point.

There is a useful algebraic description of h. The elements of $\mathrm{Möb}(\mathbb{H})$ are parametrized by the elements of

$$\mathrm{PSL}(2,\mathbb{R}) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a,b,c,d \in \mathbb{R}, ad - bc = 1 \right\} / \left\{ \pm \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

We see that the map $\omega \mapsto$ coefficient matrix of φ_ω is bijection $T^1\mathbb{H} \to \mathrm{PSL}(2,\mathbb{R})$. Applying this identification to (2.2), we obtain a conjugacy between the horocycle flow on $T^1\mathbb{H}$ and the matrix flow $h : \mathrm{PSL}(2,\mathbb{R}) \to \mathrm{PSL}(2,\mathbb{R})$

$$h^t : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

This extends to other hyperbolic surfaces. The Killing–Hopf Theorem says that any complete orientable connected hyperbolic surface M is isometric to an orbit space $\Gamma \setminus \mathbb{H}$, where Γ is a discrete subgroup of $\mathrm{Möb}(\mathbb{H})$ without elements of finite order ("torsion free"). Γ is called a *uniform lattice* if $\Gamma \setminus \mathbb{H}$ is compact, a *lattice* if $\Gamma \setminus \mathbb{H}$ has finite area, and *geometrically finite* if $\Gamma \setminus \mathbb{H}$ has finite genus. Every uniform lattice is a lattice, and every lattice is geometrically finite [Kat].

The identifications $T^1\mathbb{H} \simeq \mathrm{Möb}(\mathbb{H}) \simeq \mathrm{PSL}(2,\mathbb{R})$ turn the horocycle flow on T^1M into the matrix flow $h : \Gamma \setminus \mathrm{PSL}(2,\mathbb{R}) \to \Gamma \setminus \mathrm{PSL}(2,\mathbb{R})$

$$h^t : \Gamma \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \Gamma \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

Let m_0 denote the Riemannian volume measure on T^1M. We can use the algebraic representation of h to relate m_0 to the Haar measure of $\mathrm{PSL}(2,\mathbb{R})$, and to deduce its h–invariance. It is enough to treat the case $M = \mathbb{H}$, the general case follows from the representation $M = \Gamma \setminus \mathbb{H}$. The identification $\omega \mapsto$the coefficient matrix of φ_ω conjugates the action of $\mathrm{Möb}(\mathbb{H})$ on $T^1\mathbb{H}$ to the action

of $\mathrm{PSL}(2,\mathbb{R})$ on itself by multiplication on the left. Isometries preserve volume, so m_0 must be mapped to the left Haar measure on $\mathrm{PSL}(2,\mathbb{R})$. $\mathrm{PSL}(2,\mathbb{R})$ is unimodular: its left Haar measure is invariant under multiplication on the right. Since h acts by multiplication on the right, m_0 is h–invariant.

Theorem 2.1 (Kaimanovich). *m_0 is h–ergodic iff every bounded harmonic function on M is constant ("Liouville property").*

This is in [Kai] (see also [Su], part II).

2.2. Horocycle flows on hyperbolic surfaces with finite genus. Henceforth, unless stated otherwise, a "hyperbolic surface" means $\Gamma\setminus\mathbb{H}$, where Γ is a discrete torsion free subgroup of $\mathrm{Möb}(\mathbb{H})$.

Recall the following chain of inclusions for hyperbolic surfaces [Kat]: compact $\subset$ finite area $\subset$ finite genus. The study of measure rigidity for horocycle flows starts with the following fundamental result [F1]:

Theorem 2.2 (Furstenberg). *If M is compact, then $h : T^1M \to T^1M$ is uniquely ergodic. The invariant measure is, up to scaling, m_0.*

A non–compact hyperbolic surface of finite area has "cusps" (Fig. 1a): pieces which are isometric to $C := \langle z\mapsto z+1\rangle \setminus \{z\in\mathbb{H} : \mathrm{Im}\, z \geq a\}$ (where $a>0$). Cusps contain periodic horocycles. In fact any unit tangent vector $\omega\in T^1C$ which points north is h–periodic, and the Lebesgue measure on its orbit is a finite invariant measure. It follows that the horocycle flow is not uniquely ergodic. But it is uniquely ergodic in the broad sense:

Theorem 2.3 (Dani–Smillie). *Suppose M is a hyperbolic surface of finite area.*

1. *The ergodic invariant Radon measures are up to scaling the volume measure m_0, and the measures supported on periodic horocycles.*
2. *Every $\omega\in T^1M$ whose horocycle is not periodic is generic for m_0.*

Part 1 is in [Da], part 2 is in [DS].

We see that in the finite area case all invariant measures are finite. For infinite area surfaces there are no finite invariant measures at all, other than measures supported on periodic horocycles (Ratner [Rat1]). We discuss the finite genus case. To avoid trivial exceptions we always assume that the area is infinite, and we only consider non–elementary surfaces, i.e. surfaces $M=\Gamma\setminus\mathbb{H}$ for which Γ is not generated by a single element.

Such surfaces have "funnels". These are subsets which are isometric to $F := \langle z\mapsto \lambda z\rangle \setminus \{z\in\mathbb{H} : \mathrm{Re}\, z\geq 0\}$, where $\lambda>1$ (Fig. 1a). Funnels contain exceptional orbits: if the geodesic of $\omega\in T^1F$ tends to some $p\in\{z\in\partial\mathbb{H} : \mathrm{Re}\, z>0\}$, then $\int_0^\infty \delta_{h^t\omega}dt$ is a Radon measure. The Radon property is because the horocycle eventually enters one fundamental domain of $\langle z\mapsto\lambda z\rangle$ and stays there without accumulating anywhere.

The set of exceptional ω's constructed above is an h invariant set of positive volume. Its complement also has positive volume. It follows that m_0 is not ergodic.

There does exist an h–ergodic invariant Radon measure μ which gives any single orbit measure zero [Bu]. We describe it.

It is convenient to work in the unit disc model $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$ together with the metric $2\sqrt{dx^2 + dy^2}/(1 - x^2 - y^2)$. The map $\vartheta : \mathbb{H} \to \mathbb{D}$, $\vartheta(z) = \frac{i-z}{i+z}$ is an isometry from $\mathbb{H}$ to $\mathbb{D}$. It can be used to represent M in the form $\Gamma_{\mathbb{D}} \setminus \text{Möb}(\mathbb{D})$, where $\Gamma_{\mathbb{D}} = \vartheta\Gamma\vartheta^{-1}$. We abuse notation and write $\Gamma = \Gamma_{\mathbb{D}}$.

$T^1(\mathbb{D})$ can be identified with $\partial\mathbb{D} \times \mathbb{R} \times \mathbb{R}$ via $(e^{i\theta}, s, t) \leftrightarrow (h^t \circ g^s)(\omega(e^{i\theta}))$, where h is the horocycle flow, g is the geodesic flow, and $\omega(e^{i\theta})$ is the element of $T^1(\mathbb{D})$ based at the origin, and pointing at $e^{i\theta}$. (These are "KAN–coordinates" for $T^1(\mathbb{D}) \cong T^1\mathbb{H} \cong \text{PSL}(2, \mathbb{R})$.) In these coordinates, Γ acts by

$$\varphi_* : (e^{i\theta}, s, t) \mapsto (\varphi(e^{i\theta}), s - \log|\varphi'(e^{i\theta})|, t + a(\varphi, e^{i\theta}, s)) \quad (\varphi \in \Gamma) \tag{2.3}$$

where $a(\varphi, e^{i\theta}, s)$ is some function which does not depend on t. The horocycle flow is just the linear translation on the t–coordinate.

We continue to assume that $M = \Gamma \setminus \mathbb{D}$ is non–elementary, and let $\Lambda(\Gamma)$ denote the *limit set* of Γ, equal by definition to $\partial\mathbb{D} \cap \overline{\{\Gamma z\}}$ for some (hence all [Kat]) $z \in \mathbb{D}$. Let $\delta(\Gamma)$ denote the *critical exponent* of Γ, equal by definition to the infimum of all δ s.t. $\sum_{\varphi\in\Gamma} \exp[-\delta \operatorname{dist}(z, \varphi(z))] < \infty$. The following is in [Pat]:

Theorem 2.4 (Patterson). *There exists a probability measure ν on $\Lambda(\Gamma) \subseteq \partial\mathbb{D}$ such that $\frac{d\nu\circ\varphi}{d\nu} = |\varphi'|^{\delta(\Gamma)}$ for all $\varphi \in \Gamma$.*

One can now use (2.3) to verify by direct calculation that

$$d\mu(e^{i\theta}, s, t) := e^{\delta(\Gamma)s} d\nu(e^{i\theta}) ds dt \tag{2.4}$$

is a Γ–invariant h–invariant measure on $T^1\mathbb{D}$. Γ–invariance means that μ descends to an h–invariant Radon measure on T^1M. We call the resulting measure the *Burger measure*. It is an infinite Radon measure. The following theorem implies, through the ergodic decomposition, that it is ergodic.

Theorem 2.5 (Burger – Roblin). *Suppose $M = \Gamma \setminus \mathbb{H}$ is a non–elementary hyperbolic surface with finite genus and infinite area. The h–ergodic invariant Radon measures are up to scaling*

1. *The Burger measure;*
2. *Infinite measures carried by horocycles of unit tangent vectors whose forward geodesics escape to infinity through a funnel;*
3. *Finite measures carried by periodic horocycles whose forward geodesics escape to infinity through a cusp.*

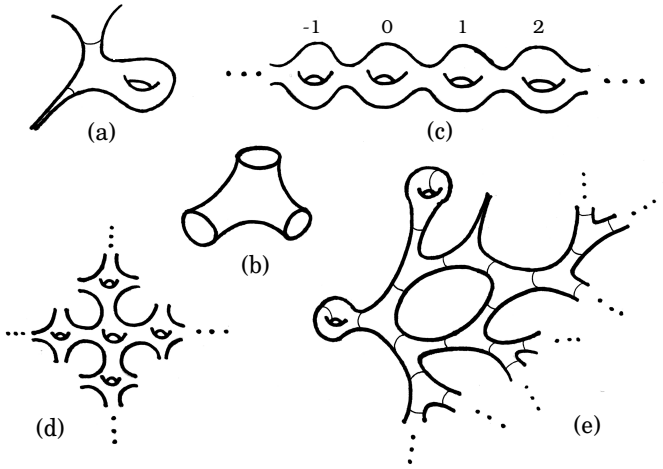


Figure 1. (a) A cusp, a funnel, and a handle; (b) A "pair of pants"; (c) A $\mathbb{Z}$–cover with its $\mathbb{Z}$–coordinates; (d) An F_2–cover of a compact surface; (e) A pants decomposition of a tame surface

The theorem was proved by Burger under the additional assumption that $\delta(\Gamma) > \frac{1}{2}$ and that M has no cusps [Bu]. The general case was done by Roblin [Ro], who also discusses extensions to variable negative curvature.

Theorem 2.6 (Schapira). *Suppose M is a hyperbolic surface of finite genus and infinite area, and let $\omega \in T^1M$. Either ω is h–periodic, or ω is exceptional, or ω is generic for the Burger measure.*

For a characterization of the generic $\omega \in T^1M$ in terms of the endpoints of their geodesics, see [Scha1], [Scha2]. For other equidistribution results which involve Burger's measure, see [Oh].

Together, theorems 2.5 and 2.6 say that the horocycle flow on a complete connected orientable hyperbolic surface of finite genus is uniquely ergodic in the broad sense.

2.3. Invariant measures in infinite genus. Horocycle flows on hyperbolic surfaces of infinite genus are not always uniquely ergodic in the broad sense, as was first discovered by Babillot and Ledrappier.

Their examples are $\mathbb{Z}^d$–covers of compact hyperbolic surfaces [BL] (Fig. 1c). These are the surfaces of the form $M = \Gamma\backslash\mathbb{H}$, where Γ is a normal subgroup of a uniform lattice Γ_0 s.t. $\Gamma_0/\Gamma \simeq \mathbb{Z}^d$. Topologically, M is a regular cover of the compact surface $M_0 = \Gamma_0\backslash\mathbb{D}$, with covering map $p(\Gamma g) = \Gamma_0 g$. The covering group

$$\mathrm{Cov}(p) := \{D : M \to M : D \text{ is a homeomorphism s.t. } p \circ D = p\}$$

is isomorphic to $\mathbb{Z}^d$.

The elements of $\mathrm{Cov}(p)$ are called "deck transformations". They are isometries, and they take the form $\Gamma z \mapsto \Gamma g_0 z$ ($g_0 \in \Gamma_0$). We parametrize the deck transformations by $D_{\underline{\xi}}$ ($\underline{\xi} \in \mathbb{Z}^d$) in such a way that $D_{\underline{\xi}+\underline{\eta}} = D_{\underline{\xi}} \circ D_{\underline{\eta}}$. The deck transformations act on T^1M by their differentials. Abusing notation, we denote this action again by $D_{\underline{\xi}}$.

Theorem 2.7 (Babillot & Ledrappier). *For each $\underline{a} \in \mathbb{R}^d$ there exists up to scaling a unique h–ergodic invariant Radon measure $m_{\underline{a}}$ s.t. $m_{\underline{a}} \circ D_{\underline{\xi}} = e^{\langle \underline{a}, \underline{\xi} \rangle} m_{\underline{a}}$ ($\underline{\xi} \in \mathbb{Z}^d$).*

The parameter $\underline{a} = \underline{0}$ corresponds to m_0, the measure induced by the Haar measure. The measures $m_{\underline{a}}$ with $\underline{a} \neq \underline{0}$ are singular. Each is infinite, globally supported, and quasi–invariant under the geodesic flow $g : T^1M \to T^1M$: $\exists c(\underline{a})$ s.t. $m_{\underline{a}} \circ g^s = e^{c(\underline{a})s} m$. For a related result on nilpotent regular covers of compact hyperbolic surfaces, see Babillot [Ba].

Theorem 2.8 (S.). *Every h–ergodic invariant Radon measure is proportional to $m_{\underline{a}}$ for some $\underline{a}$.*

See [Sa2]. Notice that although there is more than one non–trivial ergodic invariant Radon measure, the collection of these measures is still small enough to be completely described.

Babillot noticed a striking similarity between the list $\{m_{\underline{a}} : \underline{a} \in \mathbb{R}^d\}$, and the list of minimal positive eigenfunctions of the Laplacian on M [Ba]. Some definitions:

- The *hyperbolic Laplacian of* $\mathbb{H}$ is a second order differential operator on $C^2(\mathbb{H})$ s.t. $\Delta_{\mathbb{H}}(f \circ \varphi) = (\Delta_{\mathbb{H}} f) \circ \varphi$ for all $\varphi \in \mathrm{M\ddot{o}b}(\mathbb{H})$. This determines $\Delta_{\mathbb{H}}$ up to a constant. With a particular choice of constant, $\Delta_{\mathbb{H}} = y^2(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2})$.

- The *hyperbolic laplacian of* $M = \Gamma \backslash \mathbb{H}$ is $(\Delta_M f)(\Gamma z) := (\Delta_{\mathbb{H}} \widetilde{f})(z)$ where $\widetilde{f}(z) := f(\Gamma z)$. The definition is proper, because of the commutation relation between $\Delta_{\mathbb{H}}$ and $\mathrm{M\ddot{o}b}(\mathbb{H})$.

- The *positive λ–eigenfunctions* of Δ_M are the positive $F \in C^2(M)$ for which $\Delta_M F = \lambda F$. (We allow infinite L^2 norm.) We say that F is *minimal*, if $\Delta_M G = \lambda G$, $0 \leq G \leq F \Rightarrow \exists c$ s.t. $G = cF$. The minimal positive λ–eigenfunctions are the extremal rays of the cone of positive λ–eigenfunctions.

The minimal positive eigenfunctions of the Laplacian on a $\mathbb{Z}^d$–cover of a compact hyperbolic surface can be parametrized, up to scaling, by $\{F_{\underline{a}} : \underline{a} \in \mathbb{R}^d\}$, where $F_{\underline{a}} \circ D_{\underline{\xi}} = e^{\langle \underline{a}, \underline{\xi} \rangle} F_{\underline{a}}$ ($\underline{\xi} \in \mathbb{Z}^d$) (see [LP] and references therein). The similarity with the list of ergodic invariant Radon measures is obvious.

Motivated by this observation and Sullivan's work on the geodesic flow, Babillot proposed a method for getting invariant Radon measures out of positive

eigenfunctions, and conjectured that at least in some cases her method provides a bijection between the two collections. We describe Babillot's construction.

Again, it is convenient to represent $M = \Gamma \setminus \mathbb{D}$, where Γ is a discrete torsion free subgroup of Möb($\mathbb{D}$). The hyperbolic laplacian on $\mathbb{D}$ is $\Delta_{\mathbb{D}} f := [\Delta_{\mathbb{H}}(f \circ \vartheta)] \circ \vartheta^{-1}$, where $\vartheta : \mathbb{H} \to \mathbb{D}$ is the isometry $z \mapsto \frac{i-z}{i+z}$. The reader can check that $\Delta_{\mathbb{D}}$ commutes with Möb($\mathbb{D}$), and that $\Delta_{\mathbb{D}} = \frac{1}{4}(1-|z|^2)^2(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2})$.

Any positive eigenfunction of Δ_M lifts to a Γ–invariant positive eigenfunction of $\Delta_{\mathbb{D}}$. Any positive eigenfunction of $\Delta_{\mathbb{D}}$ can be represented in the form

$$F(z) = \int_{\partial \mathbb{D}} P(e^{i\theta}, z)^{\alpha} d\nu_F(e^{i\theta}),$$

where ν is a finite positive measure on $\partial\mathbb{D}$, $P(e^{i\theta}, z) = (1-|z|^2)/|e^{i\theta} - z|^2$ is Poisson's kernel, and $\alpha \geq 1/2$ (Karpelevich [Kar], see also [GJT]). If $\delta(\Gamma) \geq \frac{1}{2}$, then this representation is unique, and the Γ–invariance of F translates to the following condition on ν:

$$\frac{d\nu_F \circ \varphi}{d\nu_F} = |\varphi'|^{\alpha} \text{ for all } \varphi \in \Gamma.$$

Comparing this with (2.3), we see that the measure

$$dm_F = e^{\alpha s} d\nu_F(e^{i\theta}) ds dt \tag{2.5}$$

is a Γ–invariant, h–invariant measure on $T^1(\mathbb{D})$. Its restriction to a fundamental domain of Γ induces an h–invariant measure on $M = \Gamma \setminus \mathbb{D}$.

Thus a positive eigenfunction F gives rise to a horocycle invariant Radon measure m_F. Babillot has conjectured – in the case of infinite regular covers of compact surfaces with nilpotent covering group – that every invariant Radon measure arises this way, and that minimal eigenfunctions F lead to ergodic invariant Radon measures m_F [Ba].

Babillot's conjecture was proved for all infinite regular covers in [LS2] (see [L] for a related result in higher dimension), and later for all *tame surfaces* [Sa1]. To explain what these are, we recall some definitions and facts [Hub]:

- A hyperbolic surface with boundary is called a *pair of pants*, if it is homeomorphic to a sphere minus three disjoint open discs or points (Fig. 1b).
- Every pair of pants has three boundary components of lengths $0 \leq \ell_i < \infty$ $(i = 1, 2, 3)$, where $\ell = 0$ corresponds to a cusp. Two pairs of pants with the same triplet of lengths are isometric.
- The *norm* of a pair of pants Y is the sum of the lengths of its boundary components, and is denoted by $\|Y\|$.
- A discrete subgroup $\Gamma \subset$ Möb($\mathbb{D}$) is called a *fuchsian group*. A fuchsian group is said to be *of the first kind* if its limit set $\Lambda(\Gamma)$ equals $\partial\mathbb{D}$.

- A torsion free fuchsian group Γ is of the first kind iff $\Gamma\backslash\mathbb{D}$ can be partitioned into a countable collection of pants $\{Y_i\}$ which meet at boundary components of the same length (see e.g. [Hub]).

We call $\{Y_i\}$ a *pants decomposition* of M.

Definition 2.1. *The surface* $\Gamma \setminus \mathbb{D}$ *is called* tame*, if it admits a pants decomposition* $\{Y_i\}$ *such that* $\sup \|Y_i\| < \infty$.

It can be shown that in this case $\delta(\Gamma) \geq \frac{1}{2}$ [Sa1].

Any regular cover of a compact hyperbolic surface is tame, because it admits an infinite pants decomposition whose components fall into finitely many isometry classes. There are many other examples: if one glues a finite or countable collection of pants of bounded norm one to another in such a way that every boundary component is glued to some other boundary component of the same length and orientation, then the result is a tame complete hyperbolic surface (Fig. 1e).

We need a couple more definitions to state the result.

- A horocycle ergodic invariant Radon measure is called *trivial* if it is supported on a single horocycle made of unit tangent vectors whose forward geodesics tend to a cusp.
- A Möbius function $\varphi \in \mathrm{Möb}(\mathbb{D})$ is called *parabolic* if it has exactly one fixed point on $\partial\mathbb{D}$. A positive eigenfunction is called *trivial*, if it is of the form
$$F(z) := \sum_{g \in \Gamma/\mathrm{stab}_\Gamma(e^{i\theta})} P(g(e^{i\theta}), z)^\alpha$$
where $P(\cdot,\cdot)$ is Poisson's kernel, $e^{i\theta}$ is a fixed point of some parabolic element of Γ, and $\mathrm{stab}_\Gamma(e^{i\theta}) = \{g \in \Gamma : g(e^{i\theta}) = e^{i\theta}\}$.

The following theorem is proved in [Sa1], under slightly weaker assumptions.

Theorem 2.9 (S.). *If* $\Gamma \setminus \mathbb{D}$ *is tame, then the following map is a bijection between the non-trivial positive minimal eigenfunctions of the Laplacian on* $\Gamma\backslash\mathbb{D}$ *and the non-trivial horocycle ergodic invariant Radon measures on* $T^1(\Gamma \setminus \mathbb{D})$*:*

$$\left[F(\Gamma z) = \int_{\partial\mathbb{D}} P(e^{i\theta}, z)^\alpha d\nu(e^{i\theta})\right] \leftrightarrow \left[\begin{array}{l}\textit{The restriction of } dm = e^{\alpha s} d\nu(e^{i\theta}) ds dt \\ \textit{to a fundamental domain of } \Gamma\end{array}\right].$$

We illustrate the result by examples [LS2]:

1. *Furstenberg's Theorem:* All positive eigenfunctions on a compact surface are constant. The constant function maps to m_0. Consequently all ergodic invariant Radon measures are proportional to m_0.
2. *Dani's Theorem:* The minimal positive eigenfunctions on a hyperbolic surface of finite area are either constant, or trivial (Eisenstein series associated to cusps). So the ergodic invariant Radon measures are m_0 and trivial measures.

3. *Periodic surfaces of polynomial growth:* These are regular covers of compact hyperbolic surfaces with the property that the area of concentric balls of radius R is $O(R^\delta)$ for some δ as $R \to \infty$ (e.g. $\mathbb{Z}^d$–covers). Using Gromov's characterization of virtually nilpotent groups, it can be shown that the group of deck transformations contains a nilpotent normal subgroup N of finite index. The minimal positive eigenfunctions form a family

$$\{cF_\varphi : c > 0, \varphi : N \to \mathbb{R} \text{ a homomorphism}\},$$

where $F_\varphi \circ D = e^{\varphi(D)} F_\varphi$ for all $D \in N$ ([LP],[CG], see also [LS2]). Consequently the ergodic invariant Radon measures of the horocycle flow are

$$\{cm_\varphi : c > 0, \varphi : N \to \mathbb{R} \text{ a homomorphism}\},$$

where $m_\varphi \circ D = e^{\varphi(D)} m_\varphi$ for all $D \in N$.

There are periodic surfaces of exponential growth for which there are locally finite ergodic invariant measures which are not quasi–invariant with respect to some deck transformations, see [LS2].

Question 1. Does there exist an example of a (necessarily non–tame) surface $\Gamma \setminus \mathbb{D}$ with an h–ergodic invariant Radon measure which is not carried by a single orbit, and is not quasi invariant under the geodesic flow?

Question 2. What can be said about the infinite locally finite ergodic invariant measures for general unipotent flows on a homogenous space $\Gamma \setminus G$ when Γ is not a lattice? (The finite invariant measures are known [Rat1].)

2.4. Generic points.

At present, the generic points for horocycle flows on surfaces of infinite genus are only understood in the case of $\mathbb{Z}^d$–covers of compact surfaces.

Suppose M covers a compact surface M_0 in such a way that the group of deck transformations can be put in the form $\{D_{\underline{\xi}} : \underline{\xi} \in \mathbb{Z}^d\}$, where $D_{\underline{\xi}+\underline{\eta}} = D_{\underline{\xi}} \circ D_{\underline{\eta}}$. Choose some connected fundamental domain $\widetilde{M_0}$ for the action of the group of deck transformations on T^1M. Define the $\mathbb{Z}^d$–*coordinate* of $\omega \in T^1M$ to be the unique $\xi(\omega) \in \mathbb{Z}^d$ such that $\omega \in D_{\xi(\omega)}[\widetilde{M_0}]$ (Fig. 1c).

There is an analogy between the paths of the geodesic flow and the paths of a random walk on $\mathbb{Z}^d$. Define the *asymptotic drift of a vector* $\omega \in T^1M$ to be the following limit, if it exists:

$$\Xi(\omega) := \lim_{T\to\infty} \frac{1}{T} \xi(g^T\omega), \text{ where } g \text{ is the geodesic flow.}$$

Since g moves at unit speed, $\|\Xi(\omega)\|$ is uniformly bounded. Let

$$\mathfrak{C} := \text{closed convex hull of } \{\Xi(\omega) : \omega \in T^1M \text{ s.t. } \Xi(\omega) \text{ exists}\} \subset \mathbb{R}^d.$$

In the previous section we parametrized the ergodic invariant Radon measures of h by the way they transform under the deck transformations. One can also parametrize them by the almost sure value of $\Xi(\cdot)$:

Theorem 2.10. *Let M be a $\mathbb{Z}^d$–cover of a compact hyperbolic surface.*

1. *For every $\Xi \in int(\mathfrak{C})$ there exists an h–ergodic invariant Radon measure m_Ξ such that $\Xi(\cdot) = \Xi$ m_ξ–a.e., and this measure is unique up to scaling.*
2. *The volume measure m_0 is proportional to $m_{\underline{0}}$.*
3. *Every h–ergodic invariant Radon measure is proportional to m_Ξ for some $\Xi \in \mathrm{int}(\mathfrak{C})$.*

See [BL], and theorem 2.8.

Theorem 2.11 (S. & Schapira). *A vector $\omega \in T^1M$ is generic for some horocycle ergodic invariant Radon measure m iff $\Xi(\omega)$ exists and $\Xi(\omega) \in int(\mathfrak{C})$. In this case $m = cm_{\Xi(\omega)}$ for some $c > 0$. In particular, ω is generic for m_0 iff $\Xi(\omega) = 0$.*

Using the hyperbolicity of the geodesic flow and a standard specification argument, it is easy to construct vectors ω for which the limit $\Xi(\omega)$ does not exist. It is not difficult to arrange for ω to have a dense (horocycle) forward orbit. Thus there are abundantly many non-exceptional $\omega \in T^1M$ which are not generic for any Radon measure. This is yet another way in which h fails to be u.e. in the broad sense.

Question 1. What are the generic points for horocycle flows on nilpotent covers of compact hyperbolic surfaces?

Question 2. Suppose $M = \Gamma \setminus \mathbb{D}$ is Liouville (cf. theorem 2.1). Is it true that $\omega \in T^1M$ is generic for m_0 whenever $\frac{1}{T} \log F(\text{base point of } g^s(\omega)) \xrightarrow[T\to\infty]{} 0$ for all positive minimal eigenfunctions F? This is the case for compact surfaces, surfaces of finite area, and $\mathbb{Z}^d$–covers of compact surfaces.

2.5. Conditional unique ergodicity.

We continue to consider the special case of $\mathbb{Z}^d$–covers of compact hyperbolic surfaces.

We saw that there are infinitely many ergodic invariant measures. It turns out that up to scaling, only one of them – the volume measure – is non pathological from the ergodic theoretic point of view, in the sense that it admits a *generalized law of large numbers* in the sense of Aaronson [A2].

We explain what this means. Suppose φ is an ergodic measure preserving flow on a non–atomic measure space $(\Omega, \mathscr{B}, \mu)$, and fix some measurable set E of finite measure. We think of t as of "time" and of E as of an "event". The times when E "happened" are encoded by the function

$$x_{E,\omega}(t) := 1_E(\varphi^t(\omega)) = \begin{cases} 1 & \varphi^t(\omega) \in E; \\ 0 & \varphi^t(\omega) \notin E. \end{cases}$$

A generalized law of large numbers is a procedure for reconstructing $\mu(E)$ from $x_{E,\omega} : [0,\infty) \to \{0,1\}$:

Definition 2.2 (Aaronson). *A generalized law of large numbers (GLLN) is a function $L : \{0,1\}^{\mathbb{R}^+} \to [0,\infty)$, $L = L[x(\cdot)]$, such that for every $E \in \mathscr{B}$ of finite measure, $L[x_{E,\omega}(\cdot)] = \mu(E)$ for μ–a.e. ω.*

For example, if the underlying measure space is a probability space, then the ergodic theorem says that the following function is a GLLN:

$$L[x(t)] := \begin{cases} \lim_{T\to\infty} \frac{1}{T}\int_0^T x(t)dt & \text{the integral and limit exist} \\ 0 & \text{otherwise.} \end{cases}$$

It is obvious how to change L to make it work when $0 < \mu(\Omega) < \infty$. But if $\mu(\Omega) = \infty$, then it is not clear how to proceed, because the ergodic theorem says that in this case $\lim_{T\to\infty} \frac{1}{T}\int_0^T 1_E(\varphi^t(\omega))dt = 0$ for every E of finite measure.

It is natural to ask whether it is possible to find $a(T) = o(T)$ so that for every $E \in \mathscr{B}$, $\lim_{T\to\infty} \frac{1}{a(T)}\int_0^T 1_E(\varphi^t(\omega))dt = \mu(E)$. This is never possible [A1]:

Theorem 2.12 (Aaronson). *Let φ be an ergodic measure preserving flow on an infinite σ–finite non–atomic measure space $(\Omega, \mathscr{B}, \mu)$. Suppose $f \in L^1$, $f > 0$. There is no $a(T) > 0$ s.t. $\frac{1}{a(T)}\int_0^T f(\varphi^t(\omega))dt$ converges a.e. to a constant $c \neq 0, \infty$.*

It is still possibile that there exists $a(T) > 0$ s.t. $\frac{1}{a(T)}\int_0^T f(\varphi^t(\omega))dt$ oscillates without converging to zero or infinity. One can hope for a summability method which forces convergence to $\int f d\mu$. Such "second order ergodic theorems" are considered in [ADF]. Here is such a theorem [LS3]:

Theorem 2.13 (Ledrappier–S.). *There exists $a(T) > 0$ s.t. for all $f \in L^1(m_0)$*

$$\lim_{N\to\infty} \frac{1}{\ln\ln N}\int_3^N \frac{1}{T\ln T}\left(\frac{1}{a(T)}\int_0^T f\circ h^s ds\right)dT = \int f dm_0 \quad m_0\text{-a.e.}$$

The corresponding GLLN is $L[x(t)] := \lim_{N\to\infty} \frac{1}{\ln\ln N}\int_3^N \frac{1}{T\ln T}\left(\frac{1}{a(T)}\int_0^T x(s)ds\right)dT$ when the limit make sense, and $L[x(t)] := 0$ otherwise.

We decribe $a(T)$. Recall the definitions of $\widetilde{M}_0$ and of the $\mathbb{Z}^d$–coordinate ξ from §2.4. We pick $\omega \in \widetilde{M}_0$ randomly according to the uniform distribution on $\widetilde{M}_0$, $m_0(\cdot\,|\widetilde{M}_0)$, and consider the random variables $\omega \mapsto \xi(g^T(\omega))$. It follows from the work of Ratner [Rat2] and Katsuda & Sunada [KS] that $\xi(g^T(\omega))/\sqrt{T}$ converges in distribution to a non–degenerate multivariate Gaussian random variable N on $\mathbb{R}^d$. If $\mathrm{Cov}(N)$ is the covariance matrix of N, and $\sigma := \sqrt[d]{|\det \mathrm{Cov}(N)|}$, then

$$a(T) = \frac{m_0(\widetilde{M}_0)}{(4\pi\sigma)^{d/2}}\frac{T}{(\ln T)^{d/2}}.$$

Theorem 2.13 also holds for $\mathbb{Z}^d$–covers of non–compact surfaces of finite area, but with different $a(T)$ [LS1].

There are no similar results for any of the other h–ergodic invariant Radon measures. The reason is tied to the following property:

Definition 2.3 (Aaronson). *An ergodic invariant measure m for a flow φ (or a map T) is called* squashable*, if there is a measurable map Q which commutes with φ (or T) such that $m \circ Q^{-1} = cm$ with $c \neq 0, 1$.*

Squashable measures do not admit GLLN's: Suppose there were a GLLN $L[\cdot]$. Choose a measurable set E of positive finite measure, and some ω s.t. $L[1_A(h^t v)] = m(A)$ for $A = E, Q^{-1}E$ and $v = \omega, Q(\omega)$. We have

$$m(E) = L[1_E(h^s Q\omega)] = L[1_E(Qh^s\omega)] = L[1_{Q^{-1}E}(h^s\omega)] = m(Q^{-1}E) = cm(E),$$

whence $c = 1$, a contradiction. Thus no GLLN can exist.

Any locally finite h–ergodic invariant Radon measure m which is not proportional to m_0 is squashable, because by theorem 2.8 such a measure satisfies $m \circ D_{\underline{\xi}} = e^{\langle \underline{a}, \underline{\xi} \rangle} m$ for some vector $\underline{a} \neq \underline{0}$ and all deck transformations $D_{\underline{\xi}}$, and all deck transformations commute with h, being isometries. As a result we obtain the following "conditional unique ergodicity" result [LS2]:

Theorem 2.14 (Ledrappier – S.). *The horocycle flow on a $\mathbb{Z}^d$–cover of a compact hyperbolic surface has, up to scaling, exactly one ergodic invariant Radon measure which admits a GLLN: the volume measure m_0.*

3. Non-compact Group Extensions of Uniquely Ergodic Transformations

3.1. Group extensions. Suppose $T : \Omega \to \Omega$ is a bimeasurable bijection on a standard measurable space $(\Omega, \mathscr{B})$. Let G be a locally compact second countable topological group with left Haar measure m_G, with $m_G(G) = 1$ when G is compact. Fix a Borel function $\varphi : \Omega \to G$.

Definition 3.1. *The* skew–product *with* base $T : \Omega \to \Omega$*, and* cocycle $\varphi : \Omega \to G$ *is the map $T_\varphi : \Omega \times G \to \Omega \times G$ given by $T_\varphi : (\omega, g) \mapsto \big(T(\omega), g\varphi(\omega)\big)$. Such maps are called* group extensions*.*

In the cases considered below, T is a homeomorphism of a topological space Ω which is either a compact metric space, or a compact metric space Ω_0 minus a countable collection of points. With such examples in mind, we call a measure m on $\Omega \times G$ *locally finite*, if $m(\Omega \times K) < \infty$ for all compact $K \subset \Omega$.

If μ is a T–invariant probability measure, then $m_0 := \mu \times m_G$ is a locally finite T_φ–invariant measure, although it is not always ergodic (e.g. when φ can

be put in the form $\varphi = u(u \circ T)^{-1}$ with u Borel). The basic measure rigidity result for compact group extensions is [P], [F2]:

Theorem 3.1 (Furstenberg – Parry). *Let T be a uniquely ergodic homeomorphism of a compact metric space Ω, G be a compact Abelian group, and $\varphi : \Omega \to G$ be continuous. T_φ is uniquely ergodic iff m_0 is T_φ–ergodic.*

If G is not compact, then there could be other measures: let $\alpha : G \to \mathbb{R}$ be a measurable homomorphism, and suppose there is a probability measure ν_α on Ω s.t. $\frac{d\nu_\alpha \circ T}{d\nu_\alpha} = \exp[\alpha \circ \varphi]$, then the measure

$$dm_\alpha(\omega, g) = e^{-\alpha(g)} d\nu_\alpha(\omega) dm_G(g) \tag{3.1}$$

is a locally finite invariant measure for T_φ, as can be verified by direct calculation. Such measures are called *Maharam measures*. Some remarks:

1. If $G = \mathbb{R}$ and $\alpha = id$, then $\varphi = \log \frac{d\nu \circ T}{d\nu}$ and T_φ is called the *Radon-Nikodym extension* of T. T_φ preserves m_α, even when T does not preserve ν. This was Maharam's original motivation [M].

2. Suppose $\alpha \equiv 0$, then ν_0 is T–invariant and $m_0 = \nu_0 \times m_G$. If G is compact, then this is the only possibility, because all measurable homomorphisms $\alpha : G \to \mathbb{R}$ are trivial.

3. Maharam measures m_α with $\alpha \not\equiv 0$ do not admit GLLN's, because they are squashable: if $Q_h : (\omega, g) \mapsto (\omega, hg)$ and $h \notin \ker \alpha$, then $Q_h \circ T_\varphi = T_\varphi \circ Q_h$ and $m_\alpha \circ Q_h = c m_\alpha$ where $c = e^{-\alpha(h)} \neq 1$.

There is an obvious generalization of Maharam's construction to skew–produts over group actions. Burger's measure (2.4) and the measures arising from Babillot's bijection (2.5) are Maharam measures for the skew–product action (2.3).

The following questions arise naturally [ANSS]: Given a u.e. T, a cocycle $\varphi : \Omega \to G$, and a measurable homomorphism $\alpha : G \to \mathbb{R}$, does the Maharam measure m_α exist, and is it unique? Is it ergodic? Is every locally finite ergodic invariant measure proportional to a Maharam measure?

The following statement comes close to saying that every locally finite ergodic invariant measure is "Maharam like", after suitable change of coordinates [Rau].

Theorem 3.2 (Raugi). *If m is a locally finite T_φ–ergodic invariant measure on $\Omega \times G$, then there are a closed subgroup $H \subset G$ and Borel function $u : \Omega \to G$ s.t.*

1. *if $\widetilde{\varphi}(x) := u(x)\varphi(x)u(Tx)^{-1}$, then $\widetilde{\varphi}(x) \in H$ for m a.e. $(x, g) \in \Omega \times G$;*

2. *if $\vartheta : (x, g) \mapsto (x, gu(x)^{-1})$, then $m \circ \vartheta^{-1}$ is a $T_{\widetilde{\varphi}}$–ergodic invariant measure supported on $\Omega \times H$, and there exists a measurable homomorphism*

$\alpha : H \to \mathbb{R}$ *and a* σ*–finite measure* ν_α *on* Ω *s.t.* $\frac{d\nu_\alpha \circ T}{d\nu_\alpha} = \exp[\alpha \circ \widetilde{\varphi}]$ *and*

$$dm \circ \vartheta^{-1}(\omega, h) = e^{-\alpha(h)} d\nu_\alpha(\omega) dm_H(h). \tag{3.2}$$

3. But in general ν_α *and* $m \circ \vartheta^{-1}$ *need not be locally finite.*

The case $G = \mathbb{R}^n \times \mathbb{Z}^m$ was done in [Sa2].

The significance of part (3) is that there is an abundance of infinite σ–finite solutions to the equation $d\nu_\alpha \circ T / d\nu_\alpha = \exp[\alpha \circ \widetilde{\varphi}]$. The challenge is to determine which of them has the property that $m = (e^{-\alpha(h)} d\nu_\alpha(\omega) dm_H(h)) \circ \vartheta$ is locally finite. In some cases, and using additional structure, one can show that $H = G$ or that u is essentially bounded (i.e. $u(\omega) \in K$ a.e. for some K compact). In such cases m is locally finite iff ν_α is finite. We discuss two examples below.

3.2. Cylinder Transformations. The first example we consider is a group extension of the irrational rotation $T_\theta : \mathbb{T} \to \mathbb{T}$, $T_\theta : \omega \mapsto (\omega + \theta) \bmod 1$, where θ is a fixed irrational number and $\mathbb{T} := \mathbb{R}/\mathbb{Z}$. The cocycle is

$$\varphi : \mathbb{T} \to \mathbb{Z}\ ,\ \varphi(\omega) := \begin{cases} 1 & 0 \le \omega < \frac{1}{2} \\ -1 & \frac{1}{2} \le \omega < 1. \end{cases}$$

Let $T_{\theta,\varphi} := (T_\theta)_\varphi$, then $T_{\theta,\varphi} : (\omega, n) \mapsto ((\theta + \alpha) \bmod 1, n + \varphi(\omega))$. Note that φ and $T_{\theta,\varphi}$ are not continuous.

The original motivation was the theory of random walks [AK]. The iterates of a G–extension T_φ are given in general by

$$T_\varphi^n(\omega, g) = \left(T^n(\omega), g\varphi(\omega)\varphi(T\omega)\cdots\varphi(T^{n-1}\omega)\right). \tag{3.3}$$

The second coordinate is a random walk on G started at g. The function φ controls the jumps, and the map $T : \Omega \to \Omega$ is the driving noise. For example, if $\Omega = \{0,1\}^{\mathbb{N}}$, T is the left shift $(T\omega)_i = \omega_{i+1}$ together with the Bernoulli $(\frac{1}{2}, \frac{1}{2})$–measure, and $\varphi : \Omega \to \mathbb{Z}$ is the function $\varphi(\omega) = (-1)^{\omega_0}$, then the second coordinate in (3.3) is the simple random walk on $\mathbb{Z}$ (started at g). The interest in the cylinder transformation $T_{\theta,\varphi}$ is that the random walk it generates is driven by a map with entropy zero. Another reason T_φ is interesting is that it appears as the Poincaré section for the linear flow on the staircase surface, see below.

The measure $m_0 := m_{\mathbb{T}} \times m_{\mathbb{Z}}$ (Lebesgue times counting measure) is an invariant Radon measure for $T_{\theta,\varphi}$. This measure is ergodic [CK], see also [Sch1]. Nakada showed that Maharam's construction yields additional locally finite ergodic invariant measures [N1], [N2]:

Theorem 3.3 (Nakada). *For every* $\theta \notin \mathbb{Q}$ *and* $\alpha \in \mathbb{R}$ *there is a unique probability measure* ν *s.t.* $\frac{d\nu \circ T_\theta}{d\nu} = \exp(\alpha\varphi)$. *The measure* $dm_\alpha(\omega, n) := e^{-\alpha\varphi(\omega)} d\nu(\omega) dm_{\mathbb{Z}}(n)$ *is a conservative ergodic invariant Radon measure for* $T_{\theta,\varphi}$.

Theorem 3.4 (Aaronson, Nakada, S., & Solomyak). *Suppose $\theta \notin \mathbb{Q}$, then every ergodic invariant Radon measure for $T_{\theta,\varphi}$ is proportional to m_α for some $\alpha \in \mathbb{R}$.*

For more complicated step functions φ, see [ANSS] and [C].

If $\alpha \neq 0$, then m_α is squashable, and therefore does not admit a GLLN. Aaronson & Keane have shown in [AK] that m_0 is not squashable. In fact it admits a GLLN. This is a particular case of the following general result [A1]:

Theorem 3.5 (Aaronson). *Let $T : \Omega \to \Omega$ be a translation on a compact metric group Ω, and suppose $\varphi : \Omega \to \mathbb{Z}^d$ is measurable. Let m_0 be the product of the Haar measures on Ω and $\mathbb{Z}^d$. If m_0 is ergodic, then m_0 admits a GLLN.*

Corollary 3.1. *Suppose θ is irrational, then up to scaling, $T_{\theta,\varphi}$ has exactly one ergodic invariant Radon measure with a GLLN: m_0.*

The GLLN presented in theorem 2.13 is *finitely observable* in the sense that the knowledge of $\{1_E(\varphi^t\omega)\}_{0\le t\le T}$ for finite T yields an approximation to $m(E)$ which tends a.s. to $m(E)$ as $T \to \infty$. The GLLN provided by the existing proof of theorem 3.5 does not seem to be finitely observable.

If we assume more on θ, then we can exhibit a finitely observable GLLN, using the theory of *rational ergodicity* [A1], [A3].

Definition 3.2 (Aaronson). *A conservative ergodic measure preserving map τ on a σ–finite measure space $(\Omega, \mathscr{B}, m)$ is called* rationally ergodic*, if there are $M > 0$ and a set $A \in \mathscr{B}$ with finite positive measure s.t. for all $n \ge 1$,*

$$\left[\int_A \left(\sum_{k=0}^{n-1} 1_A \circ \tau^k\right)^2 dm\right]^{1/2} \le M\left[\int_A \left(\sum_{k=0}^{n-1} 1_A \circ \tau^k\right) dm\right]. \qquad (3.4)$$

(The other direction to Cauchy–Schwarz.)

Rationally ergodic maps admit GLLN's. To describe them, we use the following notation for Cesàro convergence: $\operatorname*{CLim}_{k\to\infty} x_k := \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} x_k$.

Theorem 3.6 (Aaronson). *Let τ be a rationally ergodic map on the space $(\Omega, \mathscr{B}, m)$, fix some A of finite positive measure which satisfies* (3.4)*, and set*

$$a_n := \frac{1}{m(A)^2}\int_A \sum_{k=1}^{n-1} 1_A \circ \tau^k dm.$$

There are $n_k \uparrow \infty$ s.t. for every $f \in L^1$, $\operatorname*{CLim}_{k\to\infty}\left[\frac{1}{a_{n_k}}\sum_{j=0}^{n_k-1} f\circ\tau^j\right] = \int f dm$ a.e.

The sequence a_n is called the *return sequence of* τ. It is unique up to asymptotic equivalence, see [A1].

Aaronson & Keane proved in [AK] that if θ is an irrational quadratic surd, then m_0 is rationally ergodic with return sequence $a_n \asymp n/\sqrt{\log n}$ ($a_n \asymp b_n$ means $C^{-1} \leq a_n/b_n \leq C$ for some $C > 0$ and all n large enough). It follows that

Theorem 3.7. *Suppose θ is an irrational root of a quadratic polynomial with integer coefficients, then $T_{\theta,\varphi}$ has, up to scaling, a unique ergodic invariant Radon measure with a GLLN: m_0. This GLLN takes the form*

$$L[x(n)] := \begin{cases} \operatorname{CLim}_{k\to\infty} \left[\frac{1}{a_{n_k}} \sum_{j=0}^{n_k-1} x(j)\right] & \text{the limit exists} \\ 0 & \text{otherwise} \end{cases}$$

for some sequences $n_k \uparrow \infty$ and $a_n \asymp n/\sqrt{\log n}$.

Question 1. What are the generic points for m_0?

We finish our discussion of T_φ with the following nice construction due to Hubert and Weiss [HW]. Let $\{R_k\}_{k\in\mathbb{Z}}$ be the sequence of tagged rectangles $R_k := [0,2] \times [0,1] \times \{k\}$ minus the points with integer coordinates. We denote the left and right vertical sides of R_k by l_k and r_k, and the top and bottom horizontal sides by t_k, b_k. For each k,

- glue l_k to r_k by the map $(x, y; k) \mapsto (x + 2, y; k)$;
- glue the left half of t_k to the right half of b_{k-1} by the map $(x, y; k) \mapsto (x + 1, y - 1; k - 1)$;
- glue the right half of t_k to the left half of b_{k+1} by the map $(x, y; k) \mapsto (x - 1, y - 1; k + 1)$.

The result is a surface of infinite area and infinite genus, which we denote by M. Fix an angle β, and let $\varphi_\beta : M \to M$ be the flow which moves each point x at unit speed on the line with slope $\tan\beta$ passing through x, while respecting identifications (Fig. 2).

Theorem 3.8 (Hubert & Weiss). *Suppose* $\tan\beta$ *is irrational, and let $Q : M \to M$ be the map $Q(x, y; k) = (x, y; k + 1)$, then*

1. *For every $\alpha \in \mathbb{R}$ there exists up to scaling exactly one ergodic invariant Radon measure m_α such that $m_\alpha \circ Q = e^\alpha m_\alpha$;*
2. *All ergodic invariant Radon measures are of this form.*

The proof is done by first checking that the union of the horizontal sides of R_k forms a Poincaré section with the properties that the roof function is constant, and the Poincaré map is conjugate to some $T_{\theta,\varphi}$ with $\theta = \theta(\beta)$ irrational [HW].

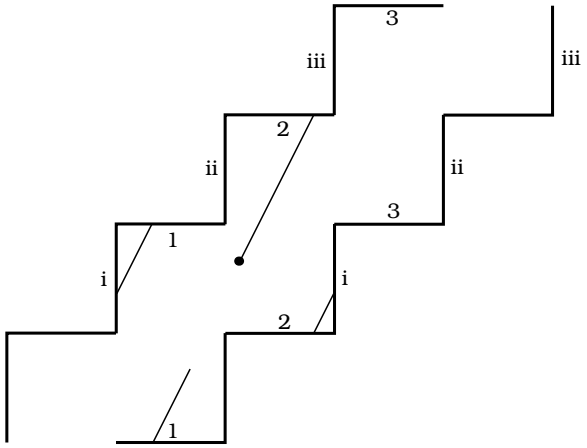


Figure 2. The linear flow on the staircase surface

Imagining other translation surfaces, one is led to the following question:

Question 2. What can be said about the locally finite ergodic invariant measures for skew products over "typical" interval exchange transformations and step function cocycles?

P. Hooper has recently obtained some very interesting related results [Hoo].

3.3. Hajian-Ito-Kakutani Maps. This example comes from the world of symbolic dynamics. Recall that the horocycle flow parametrizes the strong stable foliation of the geodesic flow: $\{h^t(\omega)\}_{t\in\mathbb{R}} = \{\omega' \in T^1M : d(g^s\omega', g^s\omega) \xrightarrow[s\to\infty]{} 0\}$. The HIK map parametrizes the symbolic dynamical analogue of the stable foliation (tail relation) for a skew–product over a subshift of finite type.

Let $\sigma : \Sigma_A^+ \to \Sigma_A^+$ be a one–sided *subshift of finite type*. This means that there is a finite set $S = \{0, \dots, N\}$ and a matrix of zeroes and ones $A = (t_{ab})_{S\times S}$ so that

$$\Sigma_A^+ := \{(x_0, x_1, \dots) \in S^{\mathbb{N}} : \forall i \geq 0,\ t_{x_i x_{i+1}} = 1\},$$
$$\text{and } \sigma : (x_0, x_1, \dots) \mapsto (x_1, x_2, \dots).$$

Endow Σ_A^+ with the metric $d(x, y) := \exp[-\min\{n \geq 0 : x_n \neq y_n\}]$. This map is *expansive*: if $d(\sigma^n x, \sigma^n y) < 1$ for all n, then $x = y$. It is topologically mixing iff there is an m s.t. all the entries of A^m are positive.

Fix some continuous function $f : \Sigma_A^+ \to \mathbb{R}^d$. The system playing the role of the geodesic flow is the (discrete time) map $\sigma_f : \Sigma_A^+ \times \mathbb{R}^d \to \Sigma_A^+ \times \mathbb{R}^d$

$$\sigma_f : (x, \xi) \mapsto (\sigma(x), \xi + f(x)).$$

We metrize $\Sigma_A^+ \times \mathbb{R}^d$ by $d((x,\xi),(y,\eta)) := d(x,y) + \|\xi - \eta\|$. One can check, using the expansivity of σ, that $d(\sigma_f^n(x,\xi), \sigma_f^n(y,\eta)) \xrightarrow[n\to\infty]{} 0$ iff

$$\exists n \text{ s.t. } \sigma^n(x) = \sigma^n(y) \text{ and } \xi - \eta = \sum_{k=0}^{\infty} [f(\sigma^k y) - f(\sigma^k x)]. \qquad (*)$$

(The sum always converges, in fact all terms with $k \geq n$ vanish.) If $(x,\xi),(y,\eta)$ satisfy $(*)$, then we write $(x,\xi) \overset{f}{\sim} (y,\eta)$. This is an equivalence relation. For an example how this equivalence relation appears as the symbolic dynamical coding of "real" foliations, see [BM] and [PoS].

Our task is to construct a map whose orbits are the equivalence classes of $\overset{f}{\sim}$. Such a map can be easily constructed using Vershik's adic transformations [V]. Here is the construction. Define $\preceq$ to be the *reverse lexicographic order* on Σ_A^+:

$$x \preceq y \Leftrightarrow \exists n \text{ s.t. } \big(x_n \leq y_n \text{ and } x_{n+k} = y_{n+k} \text{ for all } k \geq 1\big).$$

Two points x, y are $\preceq$–comparable iff $\exists n$ s.t. $\sigma^n(x) = \sigma^n(y)$. In this case we write $x \sim y$. If $x \sim y$, then there are only finitely many points between x and y (at most $|S|^n$). It follows that for all x not equivalent to a $\preceq$–maximal or minimal point, the set $\{y : y \sim x\}$ has the same order structure as $\mathbb{Z}$.

One can check that x is equivalent to a maximal (resp. minimal) point iff $\sigma^n(x)$ is maximal (resp. minimal) for some n. This leads to the following definition:

Definition 3.3 (Vershik). *Let* $\Omega := \Sigma_A^+ \setminus \{x : \exists n \, s.t. \, \sigma^n(x)$ *is maximal or minimal*$\}$. *The* adic transformation *of* Σ_A^+ *is the map* $T : \Omega \to \Omega$, $T(x) := \min\{y \in \Omega : y \succneqq x\}$.

The point is that for every $x \in \Omega$, $\{T^n(x) : n \in \mathbb{Z}\} = \{y : y \sim x\}$ for $x \in \Omega$. To get a map whose orbits are the equivalence classes of $\overset{f}{\sim}$, we make the following definition.

Definition 3.4 (Hajian–Ito–Kakutani). *Let* $f : \Sigma_A^+ \to \mathbb{R}^d$ *be a continuous function. The* HIK cocycle for f *is*

$$\varphi(x) := \sum_{k=0}^{\infty} [f(\sigma^k x) - f(\sigma^k T x)].$$

The HIK map *is* $T_\varphi : \Omega \times \mathbb{R}^d \to \Omega \times \mathbb{R}^d$, $T_\varphi : (x,\xi) \mapsto (T(x), \xi + \varphi(x))$.

A direct calculation shows that $\{T_\varphi^n(x,\xi)\}_{n\in\mathbb{Z}} = \{(y,\eta) : (y,\eta) \overset{f}{\sim} (x,\xi)\}$, and so $\{T_\varphi^n(x,\xi)\}_{n\in\mathbb{Z}} = \{(y,\eta) : d(\sigma_f^n(x,\xi), \sigma_f^n(y,\eta)) \xrightarrow[n\to\infty]{} 0\}$.

Here is an example [HIK],[AW]. Suppose $\Sigma_A^+ = S^{\mathbb{N}}$. The unique maximal point is $(N, N, N, \ldots)$, the unique minimal point is $(0, 0, 0, \ldots)$, and T is the map

$$T : (\underbrace{N, \ldots, N}_{n}, k, *) \mapsto (\underbrace{0, \ldots, 0}_{n}, k+1, *) \quad (k < N, n \geq 0) \tag{3.5}$$

Informally, T "adds one with carry to the right". Formula (3.5) makes sense for all points in $S^{\mathbb{N}} \setminus \{(1, 1, 1, \ldots)\}$. If we define $T(1, 1, 1, \ldots) := (0, 0, 0, \ldots)$, then we obtain a homeomorphism of $S^{\mathbb{N}}$, widely known under the name the *adding machine*.

Now fix some probability vector $\underline{p}_0 := (p_0, \ldots, p_N)$ on S all of whose coordinates are non–zero, let $f : \Sigma_A^+ \to \mathbb{R}$ denote the function $f(x) = -\log p_{x_0}$, and define φ to be the HIK cocycle of f. A direct calculation shows that

$$\varphi = \log\left(\frac{d\nu_0 \circ T}{d\nu_0}\right),$$

where ν_0 is the Bernoulli measure of $\underline{p}_0$. This measure is in general *not* T–invariant. But the measure $e^{-t} d\nu_0(\omega) dt$ is T_φ–invariant. Similarly, given $\alpha \in \mathbb{R}$, let $\underline{p}_\alpha$ denote the probability vector proportional to $(p_0^\alpha, \ldots, p_N^\alpha)$, and let ν_α denote the corresponding Bernoulli measure on $\Sigma_A^+ = S^{\mathbb{N}}$. Then $\log\left(\frac{d\nu_\alpha \circ T}{d\nu_\alpha}\right) = \alpha\varphi$, so $m_\alpha := e^{-\alpha t} d\nu_\alpha(\omega) dt$ is T_φ–invariant for every $\alpha \in \mathbb{R}$.

These measures are not always ergodic. The simplest example of this is when $\underline{p} = (b, \ldots, b)$ where $b = 1/|S|$. In this case φ takes values in $b\mathbb{Z}$, and the function $F(\omega, \xi) = \exp[2\pi i \xi / b]$ is T_φ–invariant. The ergodic components of m_α take the form $e^{-\alpha t} d\nu_\alpha(\omega) dm_{b\mathbb{Z}+c}$ where $m_{b\mathbb{Z}+c}$ is the counting measure on the coset $b\mathbb{Z} + c$, and $0 \leq c < b$ [AW],[HIK]. We call this phenomenon the *lattice phenomenon*.

We now turn to the case of general HIK maps, assuming only that $\sigma : \Sigma_A^+ \to \Sigma_A^+$ is topologically mixing, and that $f : \Sigma_A^+ \to \mathbb{R}^d$ has *summable variations*:

$$\sum_{n=1}^{\infty} \mathrm{var}_n f < \infty, \text{ where } \mathrm{var}_n f := \sup\{f(x) - f(y) : x_i = y_i \ (i = 0, \ldots, n-1)\}.$$

We denote $f_n := f + f \circ \sigma + \cdots + f \circ \sigma^{n-1}$. Let H_f denote the smallest closed subgroup of $\mathbb{R}^d$ which contains $\{f_n(x) - f_n(y) : \sigma^n(x) = x, \sigma^n(y) = y, n \in \mathbb{N}\}$. The following fact can be found in [Sa2] (see also [PaS])

Lemma 3.1. *There exists a function $u_f : \Sigma_A^+ \to \mathbb{R}^d$ with summable variations and a constant c_f such that $\widetilde{f} := f + u_f - u_f \circ \sigma + c_f$ takes values in H_f.*

The group H_f is invariant under addition of coboundaries and constants, so one cannot reduce the range of f further by means of a continuous coboundary.

Let φ and $\widetilde{\varphi}$ be the HIK cocycles of f and $\widetilde{f}$, respectively. Direct calculations show that $\widetilde{\varphi} = \varphi + u_f - u_f \circ T$ and that the image of $\widetilde{\varphi}$ is in H_f. The map $\vartheta(\omega, \xi) = (\omega, \xi + u_f(\omega))$ satisfies $\vartheta^{-1} \circ T_\varphi \circ \vartheta = T_{\widetilde{\varphi}}$. We see that if $H_f \neq \mathbb{R}^d$, then T_φ is conjugate to an HIK map exhibiting the lattice phenomenon.

We describe the invariant measures of T_φ.

Theorem 3.9.

1. *For every $\alpha \in \mathbb{R}^d$ there is a unique probability measure ν_α s.t. $\frac{d\nu_\alpha \circ T}{d\nu_\alpha} = e^{\langle \alpha, \varphi \rangle}$;*

2. *If $H_f = \mathbb{R}^d$, then $m_\alpha := e^{-\langle \alpha, t \rangle} d\nu_\alpha(\omega) dt$ is a T_φ–ergodic invariant locally finite measure;*

3. *If $H_f = \mathbb{R}^d$, then every T_φ–ergodic invariant locally finite measure is proportional to m_α for some $\alpha \in \mathbb{R}^d$.*

Part 1 is in [PeS], see also [ANSS]. Part 2 is because σ_f is m_α–exact [G] (see [ANSS] for details). Part 3 was proved under the assumption that f is locally constant in [ANSS] and in the general case in [Sa2].

Next we discuss the lattice case. For every $c \in \mathbb{R}^d / H_f$, let $m_{H_f + c}$ denote the measure on the coset $H_f + c$ induced by the Haar measure on H_f.

Theorem 3.10. *Suppose $H_f \neq \mathbb{R}^d$, let $\widetilde{f} := f + u_f - u_f \circ \sigma + c_f$ where u_f, c_f are given by lemma* 3.1*, and let $\widetilde{\varphi}$ denote the HIK cocycle of $\widetilde{f}$.*

1. *The locally finite ergodic invariant measures for $T_{\widetilde{\varphi}}$ are the measures proportional to $m_{\alpha,c} := e^{-\langle \alpha, t \rangle} d\nu_\alpha(\omega) dm_{H_f + c}(t)$ for some $\alpha \in \mathbb{R}^d$ and $c \in \mathbb{R}^d / H_f$.*

2. *The locally finite ergodic invariant measures for T_φ are the measures proportional to $m_{\alpha,c} \circ \vartheta$ ($\alpha \in \mathbb{R}^d, c \in \mathbb{R}^d / H_f$), where $\vartheta : (\omega, \xi) \mapsto (\omega, \xi + u_f(\omega))$.*

Theorem 3.10 was proved for $f : \Sigma_A^+ \to \mathbb{Z}^d$ s.t. $H_f = \mathbb{Z}^d$ in [ANSS], and in the general case in [Sa2].

These results show that the group H mentioned in theorem 3.2 is always equal to H_f, and that the measurable function u there can be chosen to be bounded (in fact with summable variations). Consequently the change of coordinates ϑ preserves local finiteness, and the problems mentioned in part (3) of that theorem do not arise. For examples of skew products where these problems do arise, see [Sa2],[Rau].

Finally we consider the problem of GLLN's. Here we need the stronger assumption that f is Hölder continuous. Under this assumption it is proved in [ANSS] that m_0 is rationally ergodic (cf. definition 3.2). Since all other measures are squashable, we obtain

Theorem 3.11. *If $H_f = \mathbb{R}^d$, then T_φ has, up to scaling a unique locally finite ergodic invariant measure with a GLLN: m_0. The GLLN takes the form*

$$L[x(n)] := \begin{cases} \operatorname{CLim}_{k\to\infty} \left[\frac{1}{a_{n_k}} \sum_{j=0}^{n_k-1} x(j)\right] & \text{the limit exists} \\ 0 & \text{otherwise} \end{cases}$$

for some sequences $n_k \uparrow \infty$ and $a_n \asymp n/(\log n)^{d/2}$.

For an interesting application to the study of the stable foliation for a pseudo–Anosov diffeomorphism, see [PoS].

The generic points of certain HIK maps can be described. This is ongoing work with J. Aaronson, and will be published elsewhere.

Acknowledgments

This work was partially supported by NSF grant DMS–0400687 and by the EU starting Grant ErgodicNonCompact.

References

[A1] J. Aaronson: An introduction to infinite ergodic theory. *Math. Surv. and Monog.* **50**, *American Math. Soc., Providence, RI, 1997. xii+284pp*

[A2] J. Aaronson: *The intrinsic normalizing constants of transformations preserving infinite measures*, J. d'Analyse Math. **49** (1987), 239–270.

[A3] J. Aaronson: *Rational ergodicity and a metric invariant for Markov shifts*, Israel J. Math. **27** (1977), 93–123.

[ADF] J. Aaronson, M. Denker, and A. Fisher: *Second order ergodic theorems for ergodic transformations of infinite measure spaces*, Proc. AMS **114** (1992), 115–127.

[AK] J. Aaronson and M. Keane: *The visits to zero of some deterministic random walks*, Proc. London Math. Soc. **44** (1982), 535–553.

[ANSS] J. Aaronson, H. Nakada, O. Sarig and R. Solomyak: *Invariant measures and asymptotics for some skew products*, Israel J. Math. **128** (2002), 93–134. *Corrections:* Israel J. Math. **138** (2003), 377–379.

[AW] J. Aaronson and B. Weiss: *On the asymptotics of a one–parameter family of infinite measure preserving transformations*, Bol. Soc. Brasil. Mat. (N.S.) **29** (1998), 181–193.

[Ba] M. Babillot: *On the classification of invariant measures for horospherical foliations on nilpotent covers of negatively curved manifolds.* In: *Random walks and geometry* (V.A. Kaimanovich, Ed.) de Gruyter, Berlin (2004), 319–335.

[BL] M. Babillot, F. Ledrappier: *Geodesic paths and horocycle flows on Abelian covers.* Lie groups and ergodic theory (Mumbai, 1996), 1–32, Tata Inst. Fund. Res. Stud. Math. **14** (1998), Tata Inst. Fund. Res., Bombay.

[BM] R. Bowen and B. Marcus: *Unique ergodicity of horocycle foliations*, Israel J. Math. **26** (1977), 43–67.

[Bu] M. Burger: *Horocycle flow on geometrically finite surfaces*, Duke Math. J. **61** (1990), 779–803.

[C] J.-P. Conze: *Recurrence, ergodicity and invariant measures for cocycles over rotations*, Contemp. Math. **485** (2009), 45–70.

[CG] J.-P. Conze and Y. Guivarc'h: *Propriété de droite fixe et fonctions propres des opérateurs de convolutions*, Séminaire de Probabilités, (Univ. Rennes, Rennes, 1976), Exp. No. 4, 22 pp. Dept. Math. Informat., Univ. Rennes, 1976.

[CK] J.-P. Conze and M. Keane: *Ergodicité d'un flot cylindrique*, Publ. Séminaires de Math. (Fasc. I Proba.) Rennes (1976).

[Da] S. G. Dani: *Invariant measures of horospherical flows on noncompact homogeneous spaces.* Invent. Math. **47** (1978), no. 2, 101–138.

[DS] S. G. Dani, J. Smillie: *Uniform distribution of horocycle orbits for Fuchsian groups.* Duke Math. J. **51** (1984), 185–194.

[F1] H. Furstenberg: *The unique ergodicity of the horocycle flow.* Springer Lecture Notes **318** (1972), 95–115.

[F2] H. Furstenberg: *Strict ergodicity and transformation of the torus*, Amer. J. Math. **83** (1961), 573–601.

[G] Y. Guivarc'h: *Propriétés ergodiques, en mesure infinie, de certains systémes dynamiques fibrés*, Ergodic Th. Dynam. Syst. **9** (1989), 433–453.

[GJT] Y. Guivarc'h, L. Ji, J.C. Taylor: Compactifications of symmetric spaces, *Progress in Math.* **156**, *Birkhäuser (1998), xiv+284pp.*

[HIK] A. Hajian, Y. Ito, and S, Kakutani: *Invariant measures and orbits of dissipative transformations*, Adv. Math. **9** (1972), 52–65.

[Hoo] W. P. Hooper, *personal communication.*

[Hub] J. H. Hubbard: Teichmüller Theory and applications to geometry, topology, and dynamics. Volume 1: Teichmüller theory. *xx+459 pages. Matrix Edition (2006).*

[HW] P. Hubert and B. Weiss: *Dynamics on the infinite staircase surface*, Preprint (2008).

[Kai] V. A. Kaimanovich: *Ergodic properties of the horocycle flow and classification of Fuchsian groups.* J. Dynam. Control Systems **6** (2000), no. 1, 21–56.

[Kar] F.I. Karpelevich: *The geometry of geodesics and the eigenfunctions of the laplacian on symmetric spaces,* Trans. Moskov. Math. Soc. **14**, 48–185 (1965).

[Kat] S. Katok: Fuchsian groups. *x+175 pages. Chicago Lectures in Math. The U. of Chicago Press (1992).*

[KS] A. Katsuda and T. Sunada: *Closed orbits in homology classes.* Publ. Math. IHÉS **71** (1990), 5–32.

[L] F. Ledrappier: *Invariant measures for the stable foliation on negatively curved periodic manifolds*, Ann. Inst. Fourier **58**, (2008), 85–105.

[LS1] F. Ledrappier and O. Sarig: *Fluctuations of ergodic sums for horocycle flows on $\mathbb{Z}^d$–covers of finite volume surfaces*, Disc. and Cont. Dynam. Syst. **22** (2008), 247–325.

[LS2] F. Ledrappier and O. Sarig: *Invariant measures for the horocycle flow on periodic hyperbolic surfaces*, Israel J. Math. **160** (2007), 281–315.

[LS3] F. Ledrappier and O. Sarig: *Unique ergodicity for non–uniquely ergodic horocycle flows*, Disc. and Cont. Dynam. Syst. **16** (2006), 411–433.

[LP] V. Lin and Y. Pinchover: *Manifolds with group actions and elliptic operators*, Memoirs of the AMS **112** (1994), vi+78pp.

[M] D. Maharam: *Incompressible transformations*, Fund. Math. **56** (1964), 35–50.

[N1] H. Nakada: *On a family of locally finite invariant measures for a cylinder flow*, Comment. Math. Univ. St. Paul **31** (1982), 183–189.

[N2] H. Nakada: *Piecewise linear homomorphisms of type* III *and the ergodicity of cylinder flows*, Keio Math. Sem. Rep. **7** (1982), 29–40.

[Oh] H. Oh: *Dynamics on geometrically finite hyperbolic manifolds with applications to Apollonian circle packings and beyond*. Proceedings of the International Congress of Mathematicians, India 2010

[P] W. Parry: *Compact abelian group extensions of discrete dynamical systems*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **13** (1969) 95–113.

[PaS] W, Parry and K. Schmidt: *Natural coefficients and invariants for Markov Shifts*, Invent. Math. **76** (1984), 15–32.

[Pat] S.J. Patterson: *The limit set of a Fuchsian group*. Acta Math. **136** (1976), 241–273.

[PeS] K. Petersen and K. Schmidt: *Symmetric Gibbs measures*, Trans. AMS **349** (1997), 2775–2811.

[PoS] M. Pollicott and R. Sharp: *Pseudo–Anosov foliations on periodic surfaces*, Topology and Appl. **154** (2007), 2365–2375.

[Rat1] M. Ratner: *On Raghunathan's measure conjecture.* Ann. of Math. (2) **134** (1991), no. 3, 545–607.

[Rat2] M. Ratner: *The central limit theorem for geodesic flows on n-dimensional manifolds of negative curvature.* Israel J. Math. **16** (1973), 181–197.

[Rau] A. Raugi: *Mesures invariantes ergodiques pour des produits gauches*, Bull. Soc. Math. France **135** (2007), 247–258.

[Ro] T. Roblin: *Ergodicité et équidistribution en courbure négative*, Mémoires de la Soc. Math. France **95** (2003), 1–96.

[Sa1] O. Sarig: *The horocycle flow and the Laplacian on hyperbolic surfaces of infinite genus*, Geom. Funct. Anal. **19** (2010), 1757–1821.

[Sa2] O. Sarig: *Invariant measures for the horocycle flow on Abelian covers.* Inv. Math. **157** (2004), 519–551.

[SS] O. Sarig and B. Schapira: *The generic points for the horocycle flow on a class of hyperbolic surfaces with infinite genus*, Inter. Math. Res. Not. IMRN, Vol. 2008, Article ID rnn086, 37 pages.

[Scha1] B. Schapira: *Equidistribution of the horocycles of a geometrically finite surface*. Inter. Math. Res. Notices **40** (2005), 2447–2471.

[Scha2] B. Schapira: *Density and equidistribution of one–sided horocycles of a geometrically finite hyperbolic surface*. Preprint (2009).

[Sch1] K. Schmidt: *A cylinder flow arising from irregularity of distribution*, Compositio Math. **36** (1978), 225–232.

[Sch2] K. Schmidt: *Unique ergodicity and related problems*, Ergodic Theory (Proc. Conf. Math. Forschungsint., Oberwolfach, 1978), 188–198, Lect. Notes Math. **729**, Springer, Berlin, 1979.

[Su] D. Sullivan: *Discrete conformal groups and measurable dynamics*, Bull. AMS (N.S.) **6** (1982), 57–73.

[V] A. Vershik: *A theorem on Markov approximation in ergodic theory*, Boundary value problems of mathematical physics and related questions in the theory of functions, 14. Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **115** (1982), 72–82, 306.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Richness of Chaos in the Absolute Newhouse Domain

Dmitry Turaev*

Abstract

We show that universal maps (i.e. such whose iterations approximate every possible dynamics arbitrarily well) form a residual subset in an open set in the space of smooth dynamical systems. The result implies that many dynamical systems emerging in natural applications may, on a very long time scale, have quite unexpected dynamical properties, like coexistence of many non-trivial hyperbolic attractors and repellers and attractors with all zero Lyapunov exponents. Applications to reversible and symplectic maps are also considered.

Mathematics Subject Classification (2010). Primary 37C20, 37D45; Secondary 37G25, 37J40, 37E20, 37D20, 37D25, 37D30, 37C70.

Keywords. Renormalization, homoclinic tangency, elliptic orbit, hyperbolic attractor, zero Lyapunov exponent, reversible system, Hamiltonian system

Decades of the study of dynamical systems with chaotic behavior revealed that with few exceptions these systems are more difficult than we would like. The diversity and variability of the types of chaotic dynamics occurring practically in any application are so great that nobody nowadays pursues the goal of a detailed mathematical description of the dynamics of a given system with chaos. The main source of difficulty is that the most of chaotic dynamical systems which emerge in natural applications appear to be *structurally unstable*. A system, i.e. a smooth map $f : M \to M$ of a smooth n-dimensional manifold M, or a smooth flow f_t on M, is called *structurally stable* if it is topologically equivalent to every close system (two systems are topologically equivalent if there exists a homeomorphism of M which maps the orbits of one system to the orbits of the other). A structurally unstable system is thus such that its orbit structure can be changed by an arbitrarily small (in some C^r-metric on M) perturbation.

*Mathematics Department, Imperial College, SW7 2AZ London, UK.
E-mail: dturaev@imperial.ac.uk.

Structurally unstable systems can fill open regions in the space of smooth dynamical systems. One of these regions, the so-called *Newhouse domain* $\mathcal{N}$, is the interior of the closure of the set of the systems which have a *homoclinic tangency* (that is an orbit of tangency between stable and unstable manifolds of a saddle periodic orbit). By [1, 2], this open set is non-empty, for the space of C^r-smooth maps on any manifold M of dimension $n \geq 2$ for any $r \geq 2$. Importantly, any generic family of maps which contains a map with a homoclinic tangency intersects $\mathcal{N}$ for some open set of parameters [2, 4, 14, 15, 16]. As homoclinic tangencies easily appear in a huge variety of chaotic dynamical systems for many parameter values, it follows that a great many naturally emerging models belong to the Newhouse domain for some, presumably large, regions of the parameter space[1]. Studying dynamics of maps from the Newhouse domain is therefore one of the basic questions of the mathematical chaos theory.

In [3, 5, 8, 9], it was shown, however, that by an arbitrarily small (in C^r, for any $r = 2, ..., \infty, \omega$) perturbation of any map from the Newhouse domain, one can create homoclnic tangencies of arbitrarily high orders and arbitrarily degenerate periodic points. This result shows that bifurcations of any map from $\mathcal{N}$ are too diverse, and their complete and detailed understanding is impossible. An unfolding of tangency of order m requires m parameters, and here m can be arbitrarily large, so no finite-parameter unfolding can capture all changes in the dynamics which can occur at the perturbations of a given map $f \in \mathcal{N}$. In fact, for maps from an arbitrarily small neighborhood of $f \in \mathcal{N}$ in the space of C^r-smooth maps, the relation of topological equivalence has infinitely many independent continuous invariants (in other words, for any such neighborhood the set of the equivalence classes is infinite-dimensional); the same is true if we consider weaker equivalence relations: the topological equivalence on the set of non-wandering orbits, or on the set of periodic orbits, or even on the set of stable periodic orbits [3, 5]. The goal of this paper is to describe in precise terms the scale of this variability and dynamical richness for systems from the Newhouse domain.

In my opinion, the main characteristic feature of systems from $\mathcal{N}$ is the absence of self-similarity: generically, the short-time behavior does not determine what will happen on longer time scales (contrary to Axiom A systems where a finite Markov partition determines the dynamics for all times). In order to describe this property, I use the following construction from [26, 27]. Let f be a C^r-map of an n-dimensional manifold M. Let B be any ball in M, i.e. let $B = \psi(U^n)$ where U^n is the closed unit ball in R^n and ψ is some C^r-diffeomorphism which takes U^n into M (we may take various maps ψ for the same ball B; transition from one particular choice of ψ to another corresponds

[1] to get convinced, one may take any map of a two-dimensional disc with a chaotic behavior, find a saddle periodic point and follow, numerically, its stable and unstable invariant curves; the usual picture is that, after a number of iterations, folds in the unstable curve come sufficiently close to the stable curve, so the tangencies can be created by a slight parameter tuning

to a C^r-transformation of coordinates in B). We also assume that the C^r-diffeomorphism ψ is, in fact defined on some larger ball V: $U^n \subset V \subseteq R^n$. Given positive m, the map $f^m|_B$ is a return map if $f^m(B) \cap B \neq \emptyset$. By construction, the return map $f^m|_B$ is smoothly conjugate with the map $f_{m,\psi} = \psi^{-1} \circ f^m \circ \psi$ (in order the map $f_{m,\psi}$ to be properly defined, we need to also assume that $f^m(B) \subseteq \psi(V)$). The map $f_{m,\psi}$ is a C^r-map $U^n \to R^n$, and it is solely defined by the choice of the coordinate transformation ψ and the number of iterations m (the choice of the map $\psi : U^n \to \mathcal{M}$ fixes the ball $B = \psi(U^n)$ as well). We will call the maps $f_{m,\psi}$ obtained by such procedure *renormalized iterations of* f. The set $\bigcup\limits_{m,\psi} f_{m,\psi}$ of all possible renormalized iterations of f will be called *the dynamical conjugacy class* of f. As the balls $\psi(U^n)$ can be of arbitrarily small radii, with the center situated anywhere, the dynamical conjugacy class of f captures arbitrarily fine details of the long-time behavior of f.

When we speak about dynamics of the map, we somehow describe its iterations, and the description should be insensitive to coordinate transformations. Therefore, the class of the map f, as we just have introduced it, gives some representation of the dynamics of f indeed: the larger the class, the more rich and diverse the dynamics of f is. There are some natural restrictions on this richness, as certain properties of the map f are inherited by all the maps from its class. For instance, when the topological entropy of f is zero, so is the entropy of every map from the class, any form of the hyperbolicity is inherited as well, all the maps from the class of a symplectic map are symplectic (although the symplectic form may become not a standard one), the class of a volume-contracting or a volume-preserving map contains only volume-contracting and, respectively, volume-preserving maps, an orientation-preserving map produces a class which contains orientation-preserving maps alone.

Importantly, only few of such "inheritable" properties can survive C^r-small perturbations of the map f. One of the known robust structures is the so-called *dominated splitting* (see [18]). A smooth map $f : M \to M$ of a compact n-dimensional manifold M has a dominated splitting when the tangent space at every point $x \in M$ is split into direct sum of two subspaces: $N^+(x) \oplus N^-(x) = R^n$, which depend continuously on x, which are invariant with respect to the derivative of f: $f'(x)N^+(x) = N^+(f(x))$ and $f'(x)N^-(x) = N^-(f(x))$, and which, at each $x_0 \in M$, satisfy the following requirement:

$$\lambda^-(x_0) := \overline{\lim}_{m\to+\infty} \sup_{\|u\|=1, u\in N^-(x_0)} \frac{1}{m} \ln \|f'(x_m)\cdots f'(x_0)u\| <$$
$$< \lambda^+(x_0) := \underline{\lim}_{m\to+\infty} \inf_{\|v\|=1, v\in N^+(x_0)} \frac{1}{m} \ln \|f'(x_m)\cdots f'(x_0)v\|,$$

where $x_0, x_1, \ldots, x_m, \ldots$ denotes the orbit of x_0 by f; in other words, the dominance condition means that the maximal Lyapunov exponent corresponding to the subspace $N^-(x_0)$ is strictly less than minimal Lyapunov exponent corresponding to the subspace $N^+(x_0)$. There always exist trivial splittings, with

$N^- = \emptyset$, $N^+ = R^n$ or $N^- = R^m$, $N^+ = \emptyset$. Non-trivial dominated splitting exists for uniformly hyperbolic systems (in this case $\lambda^-(x) < 0 < \lambda^+(x)$ for every x) and for uniformly partially-hyperbolic and pseudo-hyperbolic (volume-hyperbolic) systems. In general, there may be several dominated splittings for the same map, so we may have a hierarchy of subspaces $\emptyset = N_0^- \subset N_1^- \subset \cdots \subset N_k^- = R^n$, $R^n = N_0^+ \supset N_1^+ \supset \cdots \supset N_k^+ = \emptyset$ such that every pair N_j^-, N_j^+ corresponds to a dominated splitting. For any particular field $N_k^\pm(x)$ of the invariant subsets in this hierarchy, the linearized map restricted to the subset may exponentially contract (or expand) d-dimensional volumes for some $d \leq \dim N_k^\pm$. This volume contraction/expansion property is also inheritable by all renormalized iterations of f and it also persists at small smooth perturbations.

The general suspicion is that, perhaps, no other robust inheritable properties exist. This claim can be demonstrated for various examples of homoclinic bifurcations (see [24]), and can be used as a working guiding principle in the study of systems with a non-trivial dynamics:
every dynamics which is possible in U^n should be expected to occur at the bifurcations of any given n-dimensional system which has a compact invariant set without a non-trivial dominated splitting and without a volume-contraction or volume-expansion property.
This statement is not a theorem and it might be not true in some situations, still it gives a useful view on global bifurcations, as we will see in a moment.

The basic example is given by an identity map of a ball. The identity map has no kind of hyperbolic structure, neither it contracts nor expands volumes, so, according to the above stated principle, ultimately rich dynamics should be expected at the bifurcations of this map. Indeed, let us call a map f *C^r-universal* [26, 27] if its dynamical conjugacy class is C^r-dense among *all* orientation-preserving C^r-diffeomorphisms acting from the closed unit ball U^n into R^n. By the definition, the dynamics of any single universal map is ultimately complicated and rich, and the detailed understanding of it is not simpler than the understanding of all diffeomorphisms $U^n \to R^n$ altogether. At the first glance, the mere existence of C^r-univesal maps of a closed ball is not obvious for sufficiently large r. However, the following result is proven in [27].

Theorem 0.1. *For every $r = 1, \ldots, \infty$, C^r-universal diffeomorphisms of a given closed ball D exist arbitrarily close, in the C^r-metric, to the identity map of D.*

A way to use this result is to note that, as it follows from Theorem 0.1, every time we have a periodic orbit for which the corresponding first-return map $x \mapsto \bar{x}$ is, locally, identity:

$$\bar{x} \equiv x,$$

or coincides with identity up to flat (i.e. sufficiently high order) terms:

$$\bar{x} = x + o(\|x\|^r),$$

a C^r-small perturbation of the system can make the first-return map universal, i.e. bifurcations of this orbit can produce dynamics as complicated as it only possible for the given dimension of the phase space.

In examples below, we show how powerful this observation can be. We start with the so-called *absolute Newhouse domain* $\mathcal{A}$ in the space of C^r-smooth maps ($r \geq 2$) of any given manifold M, $\dim M \geq 2$. This domain is an open subset of the Newhouse domain such that no map from $\mathcal{A}$ has a non-trivial dominated splitting, nor it uniformly contracts or expands volumes. The set $\mathcal{A}$ can be constructed as the interior of the closure of the set of maps which have a particular type of *heteroclinic cycle*.

Namely, in the two-dimensional case the heteroclinic cycle is the union of 4 orbits: two saddle periodic orbits, p_1 and p_2, such that the saddle value at p_1 is less than 1 and at p_2 it is greater than 1, and two heteroclinic orbits, Γ_{12} and Γ_{21}, such that Γ_{12} corresponds to transverse intersection of $W^u(p_1)$ and $W^s(p_2)$ (the unstable manifold of p_1 and the stable manifold of p_2), and Γ_{21} corresponds to *tangency* between the other pair of invariant manifolds, $W^u(p_2)$ and $W^s(p_1)$. The saddle value is defined as the absolute value of the product of multipliers of the periodic orbit, i.e. it is the absolute value of the determinant of the derivative of the first-return map (if x_0 is a point of period l, then $f^l(x_0) = x_0$ and f^l is called the first-return map). Thus, if the saddle value is greater than 1, then the map f expands area near p_1, and f is area-contracting near p_1 if the saddle value is less than 1. So, no map with the heteroclinic cycle of the type we just described is uniformly area-contracting, nor area-expanding. The tangency between the stable and unstable manifolds forbids the existence of a non-trivial dominated splitting. When the map f is perturbed, the tangency may disappear, however new orbits of heteroclinic tangency may appear somewhere else, and indeed, as follows from [2, 7], maps with a heteroclinic cycle of the above described type are dense (in C^r, $r \geq 2$) in a non-empty open region in the space of C^r-smooth maps; moreover, the closure of this region contains all maps with such heteroclinic cycles. This region is our domain $\mathcal{A}$ in the two-dimensional case.

In the higher-dimensional case, where $n = \dim M > 2$, we consider heteroclinic cycles for which the saddle periodic orbits p_1 and p_2 have one-dimensional unstable manifolds, so the multipliers $\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jn}$ of the orbit p_j are such that $|\lambda_{j1}| > 1 > \max\limits_{k\geq 2} |\lambda_{jk}|$ for each $j = 1, 2$. For each of the points p_j, we order the multipliers according to the decrease in the absolute value, i.e. $|\lambda_{jk}| \geq |\lambda_{js}|$ if $k \leq s$. We assume then that λ_{12} is real, while the rest of the multipliers λ_{1k}, $k \geq 3$, go in complex-conjugate (non-real) pairs except, maybe, for the last one, λ_{1n}, which must be real if n is odd. For the multipliers λ_{2k}, $k \geq 2$, of the orbit p_2, we will allow only the last one, λ_{2n}, to be real if n is odd. As in the two-dimensional case, we also assume that $W^u(p_1)$ and $W^s(p_2)$ have a transverse intersection at the points of a heteroclinic orbit Γ_{12}, while $W^u(p_2)$ and $W^s(p_1)$ have a tangency at the points of the heteroclinic orbit Γ_{21}.

These conditions mean [24] that the map with such heteroclinic cycle cannot have a non-trivial dominated splitting. Indeed, if we have a dominated splitting, the spaces N^+ and N^- at a periodic point must be the invariant subspaces of the derivative of the first-return map at this point; moreover, for some $\bar{\lambda} > 0$, the space N^+ corresponds to the multipliers whose absolute value is greater than $\bar{\lambda}$, and N^- corresponds to the multipliers whose absolute value is less than $\bar{\lambda}$. As the multipliers λ_{2k}, $k \geq 2$, go in pairs of equal absolute value, for any non-trivial dominated splitting the dimension of the space N^+ at the points of the orbit p_2 must be odd. On the other hand, as the multipliers λ_{1k} with $k \geq 3$ also go in complex-conjugate pairs, the only possibility for the space N^+ at the points of the other periodic orbit, p_1, be odd-dimensional corresponds to $\dim N^+ = 1$. Since N^+ depends on the point continuously, $\dim N^+$ should be the same at the points of p_1 as at the points of p_2. Thus, the only possibility for a non-trivial dominated splitting occurs when at the points of the periodic orbits p_j, $j = 1, 2$, the space N^+ corresponds to the multiplier λ_{j1} (whose absolute value is greater than 1), and the space N^- corresponds to the rest of multipliers, i.e. N^+ must be tangent to $W^u(p_j)$ and N^- must be tangent to $W^s(p_j)$. By continuity, this would imply that N^+ would be tangent to $W^u(p_2)$ at every point of $W^u(p_2)$, and N^- would be tangent to $W^s(p_1)$ at every point of $W^s(p_1)$. As the manifolds $W^u(p_2)$ and $W^s(p_1)$ are not transverse at the points of the heteroclinic orbit Γ_{21}, we find that $N^+ \oplus N^- \neq R^n$, a contradiction to the definition of the dominated splitting.

Now, assume that $\left|\prod_{k=1}^{n} \lambda_{1k}\right| < 1$ and $\left|\prod_{k=1}^{n} \lambda_{2k}\right| > 1$, i.e. the map f contracts volume at the points of p_1 and expands volume at the points of p_2. So, the maps with the heteroclinic cycle that satisfies all these assumptions do not have a non-trivial dominated splitting and cannot be uniformly volume-contracting, nor volume-expanding. One can extract from [4, 14] that the C^r-closure of the set of the maps with such heteroclinic cycles has a non-empty interior, which is our absolute Newhouse domain $\mathcal{A}$ in the space of n-dimensional C^r-maps.

By the definition, for any map $f \in \mathcal{A}$, by an arbitrarily small perturbation of f a heteroclinic cycle of the type we just described can be created. Typically, the tangency between $W^u(p_2)$ and $W^s(p_1)$ at the points of the heteroclinic orbit Γ_{21} is quadratic, however, by an arbitrarily small (in C^r) perturbation, this tangency can be split in such a way that a new orbit of the heteroclinic tangency between $W^u(p_2)$ and $W^s(p_1)$ can be created, and for this new orbit the order of tangency can be *infinite* [8, 9]. This contradicts the usual logic stemming from singularity theory, where small perturbations usually lead to a decrease in the degeneracy. Here, the order of degeneracy may be increased without a bound (the price is that the new heteroclinic orbit which corresponds to the flat tangency is, in some sense, much longer than the original orbit of quadratic tangency). Importantly, by an additional, arbitrarily C^r-small perturbation of the heteroclinic cycle with the flat tangency, a *periodic spot* can be created (cf.

[9]). The periodic spot is a ball $D \subset M$ filled by periodic points, i.e. $f^l x \equiv x$ for every $x \in D$ and some l, the same for all $x \in D$. By applying Theorem 0.1 to the map $f^l|_D$, we thus find

Theorem 0.2. *For every $r = 2, \ldots, \infty$, the C^r-universal maps form a residual subset*[2] *in the absolute Newhouse domain.*

A more dramatic formulation of this result can be as follows: dynamics of a generic map from the absolute Newhouse domain $\mathcal{A}$ is beyond human comprehension. Indeed, just by the definition, every possible robust (i.e. common for an open set of maps) dynamical feature is present in each universal map as well. In particular, each universal map has an infinite set of attractors of all possible robust types, as well as an infinite set of repellers of all types. For example, as a corollary to Theorem 0.2, we obtain

Theorem 0.3. *For every $r = 2, \ldots, \infty$, a C^r-generic map $f \in \mathcal{A}$ has infinitely many uniformly hyperbolic attractors of every possible*[3] *topological type.*

Of course, every such map has all possible types of arbitrary uniformly-hyperbolic sets, i.e. not just attractors, also "saddles" and repellers. Similar to [7], one may show that the attractors and repellers are not separated (the closure of all the attractors has a non-empty intersection with the closure of all the repellers) for a generic map from $\mathcal{A}$. Indeed, we obtain the attractors/repellers from periodic spots, which are born in an arbitrarily small neighborhood of some heteroclinic cycles; in particular, some iteration of such spot comes close to the saddle periodic orbit which is a part of the heteroclinic cycle. By taking smaller and smaller neighborhoods of the heteroclinic cycle, we find that the limit of both attractors and repellers contain the same saddles. We note that this inseparability of the set of attractors from the set of repellers means that the Conley's fundamental construction of attractor-repeller pairs [19] cannot, generically, produce completely meaningful results.

The fact that generic maps may have an infinite (countable) set of attractors is known since [1] where the genericity of the maps with infinitely many coexisting stable periodic orbits ("sinks") was proven for *area-contracting* maps from the Newhouse domain. Moreover, the closure of the set of stable periodic orbits was shown to contain a non-trivial hyperbolic set. Generic inseparability of the set of "sinks" from the set of "sources" (completely unstable periodic orbits) was proven in [7] for the absolute Newhouse domain in the space of two-dimensional maps (i.e. when no area-contraction nor area-expansion property holds). Examples with coexistence of infinitely many *non-trivial* attractors (invariant tori, Lorenz-like attractors, Benedics-Carleson attractors) were built in [6, 10, 11, 12, 17]. Our results here show that attractors of *arbitrarily complicated nature* can coexist in unbounded number.

[2] i.e. a countable intersection of open and dense subsets

[3] for a map of the n-dimensional ball, $n \geq 2$

Maybe even more typical for the absolute Newhouse domain are strange attractors of a different nature, as described by the following

Theorem 0.4. *For every $r = 2, \dots, \infty, \omega$, a C^r-generic map $f \in \mathcal{A}$ has an uncountable (of the cardinality of continuum) set of weak attractors such that for each orbit in each of these attractors all Lyapunov exponents are zero.*

By the *weak attractor* we mean a compact, chain-transitive invariant set Y which is an intersection of a nested sequence of trapping neighborhoods, namely $Y = \bigcap_{i=1}^{\infty} D_i$ where $D_i \subseteq D_{i+1}$ and (the trapping property) $f(cl(D_i)) \subset int(D_i)$ [20, 25]. This definition means that even if we add a sufficiently small bounded noise to f, the forward iterations of any point in Y will forever stay in a small neighborhood of Y (in one of the trapping regions D_i). The chain-transitivity means that for an arbitrarily small level of the bounded noise there exists a "noisy" orbit of f which connects any two points in Y, i.e. the attractor Y contains no smaller attractor. The weak attractors we construct in Theorem 0.4 are the so-called solenoids, filled by *limit-periodic* orbits. Namely, there is a monotonically increasing sequence of integers k_i such that each of the sets D_i is a union of k_i disjoint balls D_{ij}, $j = 1, \dots, k_i$, and $f(clD_{ij}) \subset intD_{i,j+1 \bmod k_i}$ (hence k_{i+1} is always a multiple of k_i).

It is obvious that one can build such solenoids by a perturbation of periodic spots. The periodic spot itself is a chain-transitive set and it can be made a weak attractor (even asymptotically stable) by a small smooth perturbation. Every orbit in the periodic spot has all Lyapunov exponents zero. However, the maps with the *periodic* weak attractors with zero Lyapunov exponents are not generic (the set of the maps with periodic spots is, as we explained above, dense in $\mathcal{A}$, but it is not residual - by Kupka-Smale theorem). Therefore, we need a solenoid construction in order to achieve the C^r-genericty. Moreover, in contrast to the previous results which have been proven so far only in the smooth category, Theorem 0.4 holds in the real-analytic case ($r = \omega$) as well. There is a hope in the contemporary dynamical systems community that some kind of non-uniform hyperbolicity or partial hyperbolicity is a typical feature for the majority of systems. Theorem 0.4 shows, however, that this cannot be fully true in the absolute Newhouse domain.

We have used a particular type of heteroclinic cycles in order to describe the richness of dynamics and bifurcations in the absolute Newhouse domain. One can, however, show that maps with other types of homoclinic and heteroclinic cycles or other bifurcating orbits whose existence prevents the map from possessing a dominated splitting and from uniform contraction/expansion of volumes (see the corresponding criteria in [24]) also belong either to the absolute Newhouse domain itself, or to its boundary. One of the easiest examples is given by the so-called *reversible* maps. Given a smooth involution R (i.e. $R \circ R = id$) of the manifold M, a map $f : M \to M$ is called reversible if $f^{-1} = R \circ f \circ R$; such maps naturally appear as Poincaré maps in time-reversible

flows. Often, naturally appearing time-reversible flows are also Hamiltonian, however, *non-Hamiltonian* reversible flows are frequent too. A periodic point x of the reversible map f is called *symmetric* if $Rx = f^l x$ for some l (in other words, the set of points of the symmetric periodic orbit is invariant with respect to R). The symmetric periodic orbit is called *elliptic* if all its multipliers are simple and have absolute value 1. Obviously, the multipliers of a symmetric periodic orbit come in pairs: if λ is the multiplier, then λ^{-1} is also a multiplier. Therefore, a symmetric elliptic periodic orbit remains elliptic for an arbitrary reversible map sufficiently close (in C^1) to the original one. In other words, reversible maps with symmetric elliptic periodic orbits form an open subset in the space of all C^r-smooth reversible maps. This open subset is our absolute Newhouse domain in the reversible case, $\mathcal{A}_r$ (note that no non-trivial dominated splitting exists at the elliptic point, nor the map can contract/expand volumes *exponentially* at such point).

It is well-known [21] that dynamics near a typical symmetric elliptic point is pretty much conservative, e.g. invariant KAM-tori may exist. However, between the tori we have resonant periodic orbits, and one can show that by a perturbation, which is arbitrarily small in C^r, $r = 1, \dots, \infty$, and which keeps the map in the reversible class, arbitrarily degenerate resonant periodic orbits (hence - periodic spots) can be born from the elliptic orbit. Even if a periodic spot sequence is symmetric, it can be split into a pair of non-symmetric spot sequences (i.e. one sequence in the pair is taken into the other spot sequence by the involution R). Behavior near a *non-symmetric* periodic orbit (e.g. near a non-symmetric periodic spot sequence) of a reversible map does no longer need to be conservative-like or in any other way to differ from the general case (cf. [13]). Thus, by applying Theorem 0.1 to the non-symmetric periodic spot sequences which emerge near the symmetrtic elliptic orbit, we obtain

Theorem 0.5. *For every $r = 1, \dots, \infty$, the C^r-universal maps form a residual subset in $\mathcal{A}_r$. In particular, a C^r-generic map $f \in \mathcal{A}_r$ has infinitely many uniformly-hyperbolic attractors and uniformly-hyperbolic repellers of every possible topological type, and the closure of the attractors of each of such maps coincides with the closure of the repellers and contains all symmetric elliptic points.*

One may argue that the genericity notion we employ here is not necessarily adequate to the intuitive idea of "being typical". However, if we do not insist on having an infinite set of hyperbolic attractors and are satisfied with, say, one, the corresponding maps will be open and dense in $\mathcal{A}_r$. Since the emergence of hyperbolic theory in the 60-s, the problem of finding a uniformly-hyperbolic attractor in a system of natural origin has been actively discussed (see also a very interesting recent discovery in [22, 23]). Ironically, Theorem 0.5 offers amazingly simple while seemingly useless solution: any reversible map with elliptic point in general position possesses a hyperbolic attractor. Of course, this is hardly what we want, as such attractor does not represent the whole of dynamics and coexists with too many other, mainly unknown, objects.

In the case of a *symplectic* map f of an even-dimensional symplectic manifold M, we restrict the definition of the dynamical conjugacy class of f by including into it only those renormalized iterations $f_{m,\psi} = \psi^{-1} \circ f^m \circ \psi$ which all preserve the same given symplectic form on M (for example, when M is a two-dimensional disc with the standard symplectic form $dx \wedge dy$, ψ can be any map with a constant Jacobian). Though this requirement restricts possible choices of the maps ψ, the balls $\psi(U^n)$ can still be of arbitrarily small sizes and situated anywhere in M, so the such defined class of f still provides a description of the behavior of f on arbitrarily fine spatial scales. With this definition of the dynamical conjugacy class we call a symplectic map C^r-universal if the C^r-closure of its class contains all orientation-preserving *symplectic* C^r-diffeomorphisms acting from the closed unit ball U^n into R^n.

Exactly like in the above discussed case of reversible maps, the maps with elliptic periodic points form an open subset, $\mathcal{A}_s$, in the space of C^r-smooth symplectic maps. While most of the neighborhood of the elliptic point is filled by KAM-tori, resonant periodic orbits between the tori can be arbitrarily degenerate, and periodic spots can be born out of the elliptic orbit by an arbitrarily small smooth perturbation within the class of symplectic maps. By applying a "symplectic version" of Theorem 0.1 to these spots (see [26, 9] for the two-dimensional case) we obtain

Theorem 0.6. *For every $r = 1, \ldots, \infty$, the C^r-universal maps form a residual subset in the absolute Newhouse domain $\mathcal{A}_s$ in the space of symplectic maps.*

We, of course, do not have attractors or repellers here (as symplectic maps are volume-preserving). Note also that in the two-dimensional case the set $\mathcal{A}_s$ coincides with the usual Newhouse domain in the space of area-preserving maps, and in this case Theorem 0.6 holds true for the analytic case ($r = \omega$) as well [9, 28].

Symplectic maps appear as Poincaré maps for Hamiltonian systems restricted to a fixed energy level. Unless a special structure (uniform partial hyperbolicity) is imposed on the system, elliptic periodic orbits appear in Hamiltonian systems seemingly inevitably (e.g. they exist generically in energy levels near points of minimum of the Hamiltonian). By Theorem 0.6, dynamics near any such orbit can approximate iterations of an arbitrary symplectic map arbitrarily well. It is one of the most basic physics beliefs that the fundamental dynamical processes are described by Hamiltonian equations, the laws of nature. By Theorem 0.6, given any such process, we may record what the values of variables are at certain, arbitrarily long, discrete sequences of time values, and, for an arbitrary large set of such recordings, almost any, arbitrarily chosen Hamiltonian system (with an elliptic orbit somewhere) will reproduce all the records with an arbitrary high precision, just by an appropriate change of variables and arbitrarily fine tuning of parameters - with the only requirement that the number of degrees of freedom is determined correctly. In other words, for an arbitrary choice of the laws of nature one can still have an arbitrarily good

agreement with observation by making a right choice of variables. The point of view that the laws of nature are relative, and their choice is, to a certain extent, a matter of convenience, exists for a long time (see e.g. [29]); our results here provide an additional support to it.

References

[1] S. Newhouse, *Diffeomorphisms with infinitely many sinks*, Topology 13, 9–18 (1974).

[2] S. Newhouse, *The abundance of wild hyperbolic sets and non-smooth stable sets for diffeomorphisms*, Publ. Math. IHES 50, 101–151 (1979).

[3] S.Gonchenko, D.Turaev, L.Shilnikov, *On models with a structurally unstable Poincare homoclinic curve*, Sov. Math., Dokl. 44, 422–426 (1992).

[4] S. Gonchenko, D. Turaev, L. Shilnikov, *On the existence of Newhouse domains in a neighborhood of systems with a structurally unstable Poincare homoclinic curve (the higher-dimensional case)*, Russian Acad. Sci., Dokl., Math. 47, 268–273 (1993).

[5] S. Gonchenko, L. Shilnikov, D. Turaev, *On models with non-rough Poincare homoclinic curves*, Physica D 62, 1–14 (1993).

[6] S. Gonchenko, D. Turaev, L. Shilnikov, *Dynamical phenomena in multidimensional systems with a structurally unstable Poincare homoclinic curve*, Russian Acad. Sci., Dokl., Math. 47, 410–415 (1993).

[7] S. Gonchenko, D. Turaev, L. Shilnikov, *On Newhouse domains of two-dimensional diffeomorphisms with a structurally unstable heteroclinic cycle*, Proc. Steklov Inst. Math. 216, 76–125 (1997).

[8] S. Gonchenko, D. Turaev, L. Shilnikov, *Homoclinic tangencies of arbitrarily high orders in the Newhouse regions*, J. Math. Sci 105, 1738–1778 (2001).

[9] S. Gonchenko, D. Turaev, L. Shilnikov, *Homoclinic tangencies of arbitrarily high orders in conservative and dissipative two-dimensional maps*, Nonlinearity 20, 241–275 (2007).

[10] S. Gonchenko, L. Shilnikov, D. Turaev, *On dynamical properties of multidimensional diffeomorphisms from Newhouse regions: I*, Nonlinearity 21 (2008).

[11] S. Gonchenko, L. Shilnikov, D. Turaev, *On global bifurcations in three-dimensional diffeomorphisms leading to wild Lorenz-like attractors*, Regul. Chaotic Dyn. 14 (2009).

[12] S.V. Gonchenko, L.P. Shilnikov, O.V. Sten'kin, *On Newhouse Regions with Infinitely Many Stable and Unstable Invariant Tori*, Proc. Int. Conf. "Progress in Nonlinear Science", Vol. I, 80–102, University of Nizhny Novgorod, 2002.

[13] J.S.W. Lamb, O.V. Sten'kin, *Newhouse regions for reversible systems with infinitely many stable, unstable and elliptic periodic orbits*, Nonlinearity 17, 1217–1244 (2004).

[14] J. Palis, M. Viana, *High dimension diffeomorphisms displaying infinitely many periodic attractors*, Annals of Math. 140, 207–250 (1994).

[15] N. Romero, *Persistence of Homoclinic Tangencies in Higher Dimensions*, Erg. Th. Dyn. Syst. 15, 735–757(1995).

[16] P. Duarte, *Elliptic isles in families of area-preserving maps*, Erg. Th. Dyn. Syst. 28, 17811813 (2008).

[17] E. Colli, *Infinitely many coexisting strange attractors*, Ann. Inst. Henri Poincare. Non Linear Analysis 15, 539–579 (1998).

[18] C. Bonatti, L.J. Diaz, M. Viana, *Dynamics Beyond Uniform Hyperbolicity: A Global Geometric and Probabilistic Perspective*, Springer, 2005.

[19] C. Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Regional Conference Series in Mathematics, Vol. 38, American Mathematical Society, Providence, 1978.

[20] D. Ruelle, *Small random perturbations and the definition of attractors*, Lect. Notes Math. 1007, 663676 (1983).

[21] M.B. Sevryuk, *Reversible Systems*, Lect. Notes Math. 1211, Springer, 1987.

[22] S.P. Kuznetsov, *Example of a Physical System with a Hyperbolic Attractor of the Smale-Williams Type*, Phys. Rev. Lett. 95, 144101 (2005).

[23] S.P. Kuznetsov, A. Pikovsky, *Autonomous coupled oscillators with hyperbolic strange attractors*, Physica D 232, 87–102 (2007).

[24] D. Turaev, *On dimension of non-local bifurcational problems*, Bifurcation and Chaos 6, 919–948 (1996).

[25] D. Turaev, L. Shilnikov, *An example of a wild strange attractor*, Sb. Math. 189, 137–160 (1998).

[26] D. Turaev, *Polynomial approximations of symplectic dynamics and richness of chaos in non-hyperbolic area-preserving maps*, Nonlinearity 16, 1–13 (2003).

[27] D. Turaev, *On dynamics of maps close to identity*, preprint MPIM2006-161 (2006).

[28] V. Gelfreich, D. Turaev, *Universal dynamics in a neighbourhood of a generic elliptic periodic point*, Regul. Chaotic Dyn. (2010).

[29] H. Poincaré, *La Science et l'Hypothesè*, Flammarion, 1902.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Conservative Partially Hyperbolic Dynamics

Amie Wilkinson*

Abstract

We discuss recent progress in understanding the dynamical properties of partially hyperbolic diffeomorphisms that preserve volume. The main topics addressed are density of stable ergodicity and stable accessibility, center Lyapunov exponents, pathological foliations, rigidity, and the surprising interrelationships between these notions.

Mathematics Subject Classification (2010). Primary 37D30; Secondary 37C40.

Keywords. Partial hyperbolicity, dynamical foliations, Lyapunov exponents, rigidity.

Introduction

Here is a story, told at least in part through the exploits of one of its main characters. This character, like many a Hollywood (or Bollywood) star, has played a leading role in quite a few compelling tales; this one ultimately concerns the dynamics of partially hyperbolic diffeomorphisms.

We begin with a connected, compact, smooth surface S without boundary, of genus at least 2. The Gauss-Bonnet theorem tells us that the average curvature of any Riemannian metric on S must be negative, equal to $2\pi\chi(S)$, where $\chi(S)$ is the Euler characteristic of S. We restrict our attention to the metrics on S of everywhere negative curvature; among such metrics, there is a finite-dimensional moduli space of *hyperbolic* metrics, which have constant curvature. Up to a normalization of the curvature, each hyperbolic surface may be represented by a quotient $\mathbb{H}/\Gamma$, where $\mathbb{H}$ is the complex upper half plane with the metric $y^{-2}(dx^2 + dy^2)$, and Γ is a discrete subgroup of $\mathrm{PSL}(2, \mathbb{R})$, isomorphic

*Thanks to Christian Bonatti, Keith Burns, Jordan Ellenberg, Andy Hammerlindl, François Ledrappier, Charles Pugh, Mike Shub and Lee Wilkinson for reading earlier versions of this text and making several helpful suggestions. This work was supported by the NSF.

Department of Mathematics, Northwestern University, 2033 Sheridan Road, Evanston, IL 60208-2730, USA. E-mail: wilkinso@math.northwestern.edu.

to the fundamental group of S. More generally, any negatively curved metric on S lies in the conformal class of some hyperbolic metric, and the space of all such metrics is path connected. Throughout this story, S will be equipped with a negatively curved metric.

This negatively curved muse first caught the fancy of Jacques Hadamard in the late 1890's [39]. Among other things, Hadamard studied the properties of geodesics on S and a flow $\varphi_t \colon T^1S \to T^1S$ on the unit tangent bundle to S called the *geodesic flow*. The image of a unit vector v under the time-t map of this flow is obtained by following the unique unit-speed geodesic $\gamma_v \colon \mathbb{R} \to S$ satisfying $\dot{\gamma}_v(0) = v$ for a distance t and taking the tangent vector at that point:

$$\varphi_t(v) := \dot{\gamma}_v(t).$$

This geodesic flow, together with its close relatives, plays the starring role in the story told here.

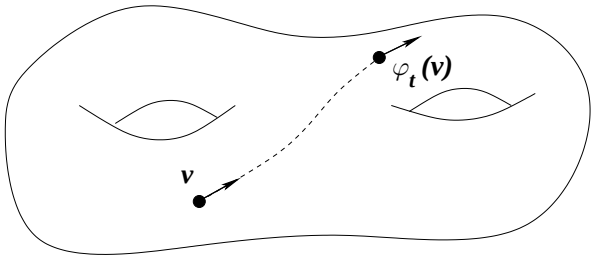


Figure 1. The geodesic flow.

A theorem of Liouville implies that φ_t preserves a natural probability measure m on T^1S, known as *Liouville measure*, which locally is just the product of normalized area on S with Lebesgue measure on the circle fibers. Poincaré recurrence then implies that almost every orbit of the geodesic flow comes back close to itself infinitely often.

In the special case where $S = \mathbb{H}/\Gamma$ is a hyperbolic surface, the unit tangent bundle T^1S is naturally identified with $PSL(2,\mathbb{R})/\Gamma$, and the action of the geodesic flow φ_t is realized by left multiplication by the diagonal matrix

$$g_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}.$$

Liouville measure is normalized Haar measure.

In his study of φ_t, Hadamard introduced the notion of the *stable manifold* of a vector $v \in T^1S$:

$$\mathcal{W}^s(v) := \left\{ w \in T^1S \mid \lim_{t\to\infty} \operatorname{dist}(\varphi_t(v), \varphi_t(w)) = 0 \right\}.$$

The proof that such sets are manifolds is a nontrivial consequence of negative curvature and a noted accomplishment of Hadamard's. Indeed, each stable

manifold $\mathcal{W}^s(v)$ is an injectively immersed, smooth copy of the real line, and taken together, the stable manifolds form a foliation $\mathcal{W}^s$ of T^1M. Similarly, one defines an *unstable manifold* by:

$$\mathcal{W}^u(v) := \left\{ w \in T^1S \mid \lim_{t\to-\infty} \operatorname{dist}(\varphi_t(v), \varphi_t(w)) = 0 \right\}$$

and denotes the corresponding unstable foliation $\mathcal{W}^u$. The foliations $\mathcal{W}^s$ and $\mathcal{W}^u$ are key supporting players in this story.

In the case where $S = \mathbb{H}/\Gamma$, the stable manifolds are orbits of the *positive horocyclic flow* on $\mathrm{PSL}(2,\mathbb{R})/\Gamma$ defined by left-multiplication by

$$h^s_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix},$$

and the unstable manifolds are orbits of the *negative horocyclic flow*, defined by left-multiplication by

$$h^u_t = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}.$$

This fact can be deduced from the explicit relations:

$$g_{-t}h^s_r g_t = h^s_{re^{-t}} \quad \text{and} \quad g_{-t}h^u_r g_t = h^u_{re^t}. \tag{1}$$

The stable and unstable foliations stratify the future and past, respectively, of the geodesic flow. It might come as no surprise that their features dictate the asymptotic behavior of the geodesic flow. For example, Hadamard obtained from the existence of these foliations and Poincaré recurrence that periodic orbits for φ_t are dense in T^1S.

Some 40 years after Hadamard received the Prix Poncelet for his work on surfaces, Eberhard Hopf introduced a simple argument that proved the ergodicity (with respect to Liouville measure) of the geodesic flow on T^1S, for any closed negatively curved surface S [44]. In particular, Hopf proved that almost every infinite geodesic in S is dense (and uniformly distributed), not only in S, but in T^1S. It was another thirty years before Hopf's result was extended by Anosov to geodesic flows for negatively curved compact manifolds in arbitrary dimension.

Up to this point the discussion is quite well-known and classical, and from here the story can take many turns. For example, for arithmetic hyperbolic surfaces, the distribution of closed orbits of the flow and associated dynamical zeta functions quickly leads us into deep questions in analytic number theory. Another path leads to the study the spectral theory of negatively curved surfaces, inverse problems and quantum unique ergodicity. The path we shall take here leads to the definition of partial hyperbolicity.

Let us fix a unit of time $t_0 > 0$ and discretize the system φ_t in these units; that is, we study the dynamics of the time-t_0 map φ_{t_0} of the geodesic flow. From a digital age perspective this is a natural thing to do; for example, to plot

the orbits of a flow, a computer evaluates the flow at discrete, usually equal, time intervals.

If we carry this computer-based analogy one step further, we discover an interesting question. Namely, a computer does not "evaluate the flow" precisely, but rather uses an *approximation to the time-t_0 map* (such as an ODE solver or symplectic integrator) to compute its orbits. To what extent does iterating this approximation retain the actual dynamical features of the flow φ_t, such as ergodicity?

To formalize this question, we consider a *diffeomorphism* $f\colon T^1S \to T^1S$ such that the C^1 distance $d_{C^1}(f, \varphi_{t_0})$ is small. Note that f in general will no longer embed in a flow. While we assume that the distance from f to φ_{t_0} is small, this is no longer the case for the distance from f^n to φ_{nt_0}, when n is large.

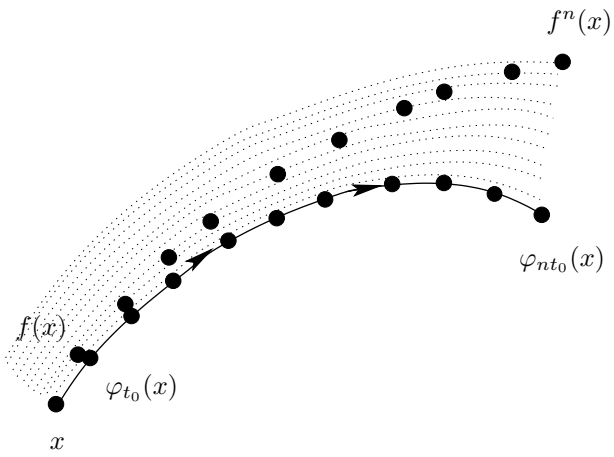


Figure 2. $f^n(x)$ is not a good approximation to $\varphi_{nt_0}(x)$.

The earliest description of the dynamics of such a perturbation f comes from a type of structural stability theorem proved by Hirsch, Pugh, and Shub [43]. The results there imply in particular that if $d_{C^1}(f, \varphi_{t_0})$ is sufficiently small, then there exists an f-invariant *center foliation* $\mathcal{W}^c = \mathcal{W}^c(f)$ that is homeomorphic to the orbit foliation $\mathcal{O}$ of φ_t. The leaves of $\mathcal{W}^c$ are smooth. Moreover, the homeomorphism $h\colon T^1S \to T^1S$ sending $\mathcal{W}^c$ to $\mathcal{O}$ is close to the identity and $\mathcal{W}^c$ is the unique such foliation.

The rest of this paper is about f and, in places, the foliation $\mathcal{W}^c(f)$.

What is known about f is now substantial, but far from complete. For example, the following basic problem is open.

Problem. Determine whether f has a dense orbit. More precisely, does there exist a neighborhood $\mathcal{U}$ of φ_{t_0} in the space $\mathrm{Diff}^r(T^1S)$ of C^r diffeomorphisms of T^1S (for some $r \geq 1$) such that every $f \in \mathcal{U}$ is topologically transitive?

Note that φ_{t_0} is ergodic with respect to volume m, and hence is itself topologically transitive. In what follows, we will explain results from the last 15 years implying that any perturbation of φ_{t_0} *that preserves volume* is ergodic, and hence has a dense orbit. For perturbations that do not preserve volume, a seminal result of Bonatti and Díaz shows that φ_{t_0} can be approximated arbitrarily well by C^1-open sets of transitive diffeomorphisms [9]. But the fundamental question of whether φ_{t_0} lives in such an open set remains unanswered.

In most of the discussion here, we will work in the conservative setting, in which the diffeomorphism f preserves a volume probability measure. To fix notation, M will always denote a connected, compact Riemannian manifold without boundary, and m will denote a probability volume on M. For $r \geq 1$, we denote by $\mathrm{Diff}^r_m(M)$ the space of C^r diffeomorphisms of M preserving m, equipped with the C^r topology.

1. Partial Hyperbolicity

The map φ_{t_0} and its perturbation f are concrete examples of partially hyperbolic diffeomorphisms. A diffeomorphism $f\colon M \to M$ of a compact Riemannian manifold M is *partially hyperbolic* if there exists an integer $k \geq 1$ and a nontrivial, Df-invariant, continuous splitting of the tangent bundle

$$TM = E^s \oplus E^c \oplus E^u$$

such that, for any $p \in M$ and unit vectors $v^s \in E^s(p)$, $v^c \in E^c(p)$, and $v^u \in E^u(p)$:

$$\begin{array}{rcccl} \|D_p f^k v^s\| & < & 1 & < & \|D_p f^k v^u\|, \quad \text{and} \\ \|D_p f^k v^s\| & < & \|D_p f^k v^c\| & < & \|D_p f^k v^u\|. \end{array}$$

Up to a change in the Riemannian metric, one can always take $k = 1$ in this definition [37]. In the case where E^c is the trivial bundle, the map f is said to be *Anosov*. The central example φ_{t_0} is partially hyperbolic: in that case, the bundle $E^c = \mathbb{R}\dot{\varphi}$ is tangent to the orbits of the flow, and E^s and E^u are tangent to the leaves of $\mathcal{W}^s$ and $\mathcal{W}^u$, respectively.

Partial hyperbolicity is a C^1-open condition: any diffeomorphism sufficiently C^1-close to a partially hyperbolic diffeomorphism is itself partially hyperbolic. Hence the perturbations of φ_{t_0} we consider are also partially hyperbolic. For an extensive discussion of examples of partially hyperbolic dynamical systems, see the survey articles [20, 41, 62] and the book [55]. Among these examples are: the frame flow for a compact manifold of negative sectional curvature and most affine transformations of compact homogeneous spaces.

As is the case with the example φ_{t_0}, the stable and unstable bundles E^s and E^u of an arbitrary partially hyperbolic diffeomorphism are always tangent to foliations, which we will again denote by $\mathcal{W}^s$ and $\mathcal{W}^u$ respectively; this is a consequence of partial hyperbolicity and a generalization of Hadamard's argument. By contrast, the center bundle E^c need not be tangent to a foliation,

and can even be nowhere integrable. In many cases of interest, however, there is also a center foliation $\mathcal{W}^c$ tangent to E^c: the content of the Hirsch-Pugh-Shub work in [43] is the properties of systems that admit such foliations, known as "normally hyperbolic foliations."

There is a natural and slightly less general notion than integrability of E^c that appears frequently in the literature. We say that a partially hyperbolic diffeomorphism $f\colon M \to M$ is *dynamically coherent* if the subbundles $E^c \oplus E^s$ and $E^c \oplus E^u$ are tangent to foliations $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$, respectively, of M. If f is dynamically coherent, then the center bundle E^c is also integrable: one obtains the center foliation $\mathcal{W}^c$ by intersecting the leaves of $\mathcal{W}^{cs}$ and $\mathcal{W}^{cu}$. The examples φ_{t_0} are dynamically coherent, as are their perturbations (by [43]: see [24] for a discussion).

2. Stable Ergodicity and the Pugh-Shub Conjectures

Brin and Pesin [16] and independently Pugh and Shub [57] first examined the ergodic properties of partially hyperbolic systems in the early 1970's. The methods they developed give an ergodicity criterion for partially hyperbolic $f \in \mathrm{Diff}^2_m(M)$ satisfying the following additional hypotheses:

(a) the bundle E^c is tangent to a C^1 foliation $\mathcal{W}^c$, and

(b) f acts isometrically (or nearly isometrically) on the leaves of $\mathcal{W}^c$.

In [16] it is shown that such an f is ergodic with respect to m if it satisfies a condition called accessibility.

Definition 2.1. A partially hyperbolic diffeomorphism $f : M \to M$ is *accessible* if any point in M can be reached from any other along an *su-path*, which is a concatenation of finitely many subpaths, each of which lies entirely in a single leaf of $\mathcal{W}^s$ or a single leaf of $\mathcal{W}^u$.

This ergodicity criterion applies to the discretized geodesic flow φ_{t_0}: its center bundle is tangent to the orbit foliation for φ_t, which is smooth, giving (a). The action of φ_{t_0} preserves the nonsingular vector field $\dot{\varphi}$, which implies (b). It is straightforward to see that if S is a hyperbolic surface, then φ_{t_0} is accessible: the stable and unstable foliations are orbits of the smooth horocyclic flows h^s_t and h^u_t, respectively, and matrix multiplication on the level of the Lie algebra $\mathfrak{sl}_2$ shows that locally these flows generate all directions in $\mathrm{PSL}(2,\mathbb{R})$:

$$\frac{1}{2}\left[\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\right] = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}; \tag{2}$$

the matrices appearing on the left are infinitesimal generators of the horocyclic flows, and the matrix on the right generates the geodesic flow. Since φ_{t_0} is accessible, it is ergodic.

Now what of a small perturbation of φ_{t_0}? As mentioned above, any $f \in \mathrm{Diff}^2_m(T^1S)$ sufficiently C^1 close to φ_{t_0} also has a center foliation $\mathcal{W}^c$, and the action of f on the leaves is nearly isometric. With some work, one can also show that f is accessible (this was carried out in [16]). There is one serious reason why the ergodicity criterion of [16] cannot be applied to f: the foliation $\mathcal{W}^c$ is not C^1. The leaves of $\mathcal{W}^c$ are C^1, and the tangent spaces to the leaves vary continuously, but they do not vary smoothly. We will explore in later sections the extent to which $\mathcal{W}^c$ fails to be smooth, but for now suffice it to say that $\mathcal{W}^c$ is pathologically bad, not only from a smooth perspective but also from a measure-theoretic one.

The extent to which $\mathcal{W}^c$ is bad was not known at the time, but there was little hope of applying the existing techniques to perturbations of φ_{t_0}. The first major breakthrough in understanding the ergodicity of perturbations of φ_{t_0} came in the 1990's:

Theorem A (Grayson-Pugh-Shub [38]). *Let S be a hyperbolic surface, and let φ_t be the geodesic flow on T^1S. Then φ_{t_0} is* stably ergodic*: there is a neighborhood $\mathcal{U}$ of φ_{t_0} in $\mathrm{Diff}^2_m(T^1S)$ such that every $f \in \mathcal{U}$ is ergodic with respect to m.*

The new technique introduced in [38] was a dynamical approach to understanding Lebesgue density points which they called *juliennes*. The results in [38] were soon generalized to the case where S has variable negative curvature [74] and to more general classes of partially hyperbolic diffeomorphisms [59, 60]. Not long after [38] appeared, Pugh and Shub had formulated an influential circle of conjectures concerning the ergodicity of partially hyperbolic systems.

Conjecture 1 (Pugh-Shub [58]). *On any compact manifold, ergodicity holds for an open and dense set of C^2 volume preserving partially hyperbolic diffeomorphisms.*

This conjecture can be split into two parts using the concept of accessibility.

Conjecture 2 (Pugh-Shub [58]). *Accessibility holds for an open and dense subset of C^2 partially hyperbolic diffeomorphisms, volume preserving or not.*

Conjecture 3 (Pugh-Shub [58]). *A partially hyperbolic C^2 volume preserving diffeomorphism with the essential accessibility property is ergodic.*

Essential accessibility is a measure-theoretic version of accessibility that is implied by accessibility: f is essentially accessible if for any two positive volume sets A and B, there exists an su-path in M connecting some point in A to some point in B – see [20] for a discussion.

In the next two sections, I will report on progress to date on these conjectures.

Further remarks.

1. Volume-preserving Anosov diffeomorphisms (where $\dim E^c = 0$) are always ergodic. This was proved by Anosov in his thesis [1]. Note that Anosov diffeomorphisms are also accessible, since in that case the foliations $\mathcal{W}^s$ and $\mathcal{W}^u$ are transverse. Hence all three conjectures hold true for Anosov diffeomorphisms.

2. It is natural to ask whether partial hyperbolicity is a necessary condition for stable ergodicity. This is true when M is 3-dimensional [27] and also in the space of symplectomorphisms [45, 68], but not in general [71]. What is true is that the related condition of having a dominated splitting is necessary for stable ergodicity (see [27]).

3. One can also ask whether for partially hyperbolic systems, stable ergodicity implies accessibility. If one works in a sufficiently high smoothness class, then this is not the case, as was shown in the groundbreaking paper of F. Rodríguez Hertz [61], who will also speak at this congress. Hertz used methods from KAM theory to find an alternate route to stable ergodicity for certain essentially accessible systems.

4. On the other hand, it is reasonable to expect that some form of accessibility is a necessary hypothesis for a general stable ergodicity criterion for partially hyperbolic maps (see the discussion at the beginning of [24]). Unlike Anosov diffeomorphisms, which are always ergodic, partially hyperbolic diffeomorphisms need not be ergodic. For example, the product of an Anosov diffeomorphism with the identity map on any manifold is partially hyperbolic, but certainly not ergodic. See also Theorem 11.16 in [10].

3. Accessibility

In general, the stable and unstable foliations of a partially hyperbolic diffeomorphism are not smooth (though they are not pathological, either – see below). Hence it is not possible in general to use infinitesimal techniques to establish accessibility the way we did in equation (2) for the discretized hyperbolic geodesic flow. The C^1 topology allows for enough flexibility in perturbations that Conjecture 2 has been completely verified in this context:

Theorem B (Dolgopyat-Wilkinson [31]). *For any $r \geq 1$, accessibility holds for a C^1 open and dense subset of the partially hyperbolic diffeomorphisms in* $\mathrm{Diff}^r(M)$*, volume-preserving or not.*

Theorem B also applies inside the space of partially hyperbolic symplectomorphisms.

More recently, the complete version of Conjecture 3 has been verified for systems with 1-dimensional center bundle.

Theorem C (Rodríguez Hertz-Rodríguez Hertz-Ures [63]). *For any $r \geq 1$, accessibility is C^1 open and C^r dense among the partially hyperbolic diffeomorphisms in* $\mathrm{Diff}^r_m(M)$ *with one-dimensional center bundle.*

This theorem was proved earlier in a much more restricted context by Niţică-Török [54]. The C^1 openness of accesssibility was shown in [28]. A version of Theorem C for non-volume preserving diffeomorphisms was later proved in [19].

The reason that it is possible to improve Theorem B from C^1 density to C^r density in this context is that the global structure of accessibility classes is well-understood. By *accessibility class* we mean an equivalence class with respect to the relation generated by *su*-paths. When the dimension of E^c is 1, accessibility classes are (C^1 immersed) submanifolds. Whether this is always true when $\dim(E^c) > 1$ is unknown and is an important obstacle to attacking the general case of Conjecture 2.

Further remarks.

1. More precise criteria for accessibility have been established for special classes of partially hyperbolic systems such as discretized Anosov flows, skew products, and low-dimensional systems [23, 21, 64].

2. Refined formulations of accessibility have been used to study higher-order statistical properties of certain partially hyperbolic systems, in particular the discretized geodesic flow [29, 52]. The precise relationship between accessibility and rate of mixing (in the absence of other hypotheses) remains a challenging problem to understand.

3. Accessibility also plays a key role in a recently developed Livsič theory for partially hyperbolic diffeomorphisms, whose conclusions closely mirror those in the Anosov setting [47, 73].

4. Ergodicity

Conjecture 1 has been verified under one additional, reasonably mild hypothesis:

Theorem D (Burns-Wilkinson [22]). *Let f be C^2, volume-preserving, partially hyperbolic and center bunched. If f is essentially accessible, then f is ergodic, and in fact has the Kolmogorov property.*

The additional hypothesis is "center bunched." A partially hyperbolic diffeomorphism f is *center bunched* if there exists an integer $k \geq 1$ such that for any $p \in M$ and any unit vectors $v^s \in E^s(p)$, $v^c, w^c \in E^c(p)$, and $v^u \in E^u(p)$:

$$\|D_p f^k v^s\| \cdot \|D_p f^k w^c\| < \|D_p f^k v^c\| < \|D_p f^k v^u\| \cdot \|D_p f^k w^c\|. \tag{3}$$

As with partial hyperbolicity, the definition of center bunching depends only on the smooth structure on M and not the Riemannian structure; if (3) holds for a

given metric and $k \geq 1$, one can always find another metric for which (3) holds with $k = 1$ [37]. In words, center bunching requires that the non-conformality of $Df \mid E^c$ be dominated by the hyperbolicity of $Df \mid E^u \oplus E^s$. Center bunching holds automatically if the restriction of Df to E^c is conformal in some metric (for this metric, one can choose $k = 1$). In particular, if E^c is one-dimensional, then f is center bunched. In the context where $\dim(E^c) = 1$, Theorem D was also shown in [63].

Combining Theorems C and D we obtain:

Corollary 1. *The Pugh-Shub conjectures hold true among the partially hyperbolic diffeomorphisms with* 1*-dimensional center bundle.*

Further remarks.

1. The proof of Theorem D builds on the original argument of Hopf for ergodicity of geodesic flows and incorporates a refined theory of the juliennes originally introduced in [38].

2. It appears that the center bunching hypothesis in Theorem D cannot be removed without a significantly new approach. On the other hand, it is possible that Conjecture 1 will yield to other methods.

3. Formulations of Conjecture 1 in the C^1 topology have been proved for low-dimensional center bundle [12, 61] and for symplectomorphisms [3]. These formulations state that ergodicity holds for a *residual* subset in the C^1 topology.

5. Exponents

By definition, a partially hyperbolic diffeomorphism produces uniform contraction and expansion in the directions tangent to E^s and E^u, respectively. In none of the results stated so far do we make any precise assumption on the growth of vectors in E^c beyond the coarse bounds that come from partial hyperbolicity and center bunching. In particular, an ergodic diffeomorphism in Theorem D can have periodic points of different indices, corresponding to places in M where E^c is uniformly expanded, contracted, or neither. The power of the julienne-based theory is that the hyperbolicity in $E^u \oplus E^s$, when combined with center bunching and accessibility, is enough to cause substantial mixing in the system, regardless of the precise features of the dynamics on E^c.

On the other hand, the asymptotic expansion/contraction rates in E^c can give additional information about the dynamics of the diffeomorphism, and is a potentially important tool for understanding partially hyperbolic diffeomorphisms that are not center bunched.

A real number λ is a *center Lyapunov exponent* of the partially hyperbolic diffeomorphism $f : M \to M$ if there exists a nonzero vector $v \in E^c$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \|Df^n(v)\| = \lambda. \tag{4}$$

If f preserves m, then Oseledec's theorem implies that the limit in (4) exists for each $v \in E^c(x)$, for m-almost every x. When the dimension of E^c is 1, the limit in (4) depends only on x, and if in addition f is ergodic with respect to m, then the limit takes a single value m-almost everywhere.

Theorem E (Shub-Wilkinson [70]). *There is an open set $\mathcal{U} \subset \mathrm{Diff}^\infty_m(\mathbb{T}^3)$ of partially hyperbolic, dynamically coherent diffeomorphisms of the 3-torus $\mathbb{T}^3 = \mathbb{R}^3/\mathbb{Z}^3$ for which:*

- *the elements of $\mathcal{U}$ approximate arbitrarily well (in the C^∞ topology) the linear automorphism of $\mathbb{T}^3$ induced by the matrix:*

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- *the elements of $\mathcal{U}$ are ergodic and have positive center exponents, m-almost everywhere.*

Note that the original automorphism A has vanishing center exponents, *everywhere* on $\mathbb{T}^3$, since A is the identity map on the third factor. Yet Theorem E says that a small perturbation mixing the unstable and center directions of A creates expansion in the center direction, almost everywhere on $\mathbb{T}^3$.

The systems in $\mathcal{U}$ enjoy the feature of being *non-uniformly hyperbolic*: the Lyapunov exponents in every direction (not just center ones) are nonzero, m-almost everywhere. The well-developed machinery of Pesin theory guarantees a certain level of chaotic behavior from nonuniform hyperbolicity alone. For example, a nonuniformly hyperbolic diffeomorphism has at most countably many ergodic components, and a mixing partially hyperbolic diffeomorphism is Bernoulli (i.e. abstractly isomorphic to a Bernoulli process). A corollary of Theorem E is that the elements of $\mathcal{U}$ are Bernoulli systems.

The constructions in [70] raise the question of whether it might be possible to "remove zero exponents" from any partially hyperbolic diffeomorphism via a small perturbation. If so, then one might be able to bypass the julienne based theory entirely and use techniques from Pesin theory instead as an approach to Conjecture 1. More generally, and wildly optimistically, one might ask whether *any* $f \in \mathrm{Diff}^2_m(M)$ with at least one nonzero Lyapunov exponent on a positive measure set might be perturbed to produce nonuniform hyperbolicity on a positive measure set (such possibilities are discussed in [70]).

There is a partial answer to these questions in the C^1 topology, due to Baraveira and Bonatti [7]. The results there imply in particular that if $f \in$

$\mathrm{Diff}^r_m(M)$ is partially hyperbolic, then there exists $g \in \mathrm{Diff}^r_m(M)$, C^1-close to f so that the *sum* of the center Lyapunov exponents is nonzero.

Further remarks.

1. Dolgopyat proved that the same type of construction as in [70] can be applied to the discretized geodesic flow φ_{t_0} for a negatively curved surface S to produce perturbations with nonzero center exponents [30]. See also [66] for further generalizations of [70].

2. An alternate approach to proving Conjecture 1 has been proposed, taking into account the center Lyapunov exponents [18]. For systems with $\dim(E^c) = 2$, this program has very recently been carried out in the C^1 topology in [65], using a novel application of the technique of blenders, a concept introduced in [9].

6. Pathology

There is a curious by-product of nonvanishing Lyapunov exponents for the open set $\mathcal{U}$ of examples in Theorem E. By [43], there is a center foliation $\mathcal{W}^c$ for each $f \in \mathcal{U}$, homeomorphic to the trivial $\mathbb{R}/\mathbb{Z}$ fibration of $\mathbb{T}^3 = \mathbb{T}^2 \times \mathbb{R}/\mathbb{Z}$; in particular, the center leaves are all compact. The almost everywhere exponential growth associated with nonzero center exponents is incompatible with the compactness of the center foliation, and so the full volume set with positive center exponent must meet almost every leaf in a zero set (in fact a finite set [67]).

The same type of phenomenon occurs in perturbations of the discretized geodesic flow φ_{t_0}. While in that case the leaves of $\mathcal{W}^c$ are mostly noncompact, they are in a sense "dynamically compact." An adaptation of the arguments in [67] shows that any perturbation of φ_{t_0} with nonvanishing center exponents, such as those constructed by Dolgopyat in [30], have atomic disintegration of volume along center leaves.

Definition 6.1. A foliation $\mathcal{F}$ of M with smooth leaves has *atomic disintegration of volume* along its leaves if there exists $A \subset M$ such that

- $m(M \setminus A) = 0$, and
- A meets each leaf of $\mathcal{F}$ in a discrete set of points (in the leaf topology).

At the opposite end of the spectrum from atomic disintegration of volume is a property called absolute continuity. A foliation $\mathcal{F}$ is *absolutely continuous* if holonomy maps between smooth transversals send zero volume sets to zero volume sets. If $\mathcal{F}$ has smooth leaves and is absolutely continuous, then for every set $A \subset M$ satisfying $m(M \setminus A) = 0$, the intersection of A with the leaf $\mathcal{F}$ through m-almost every point in M has full leafwise Riemannian volume. In this sense Fubini's theorem holds for absolutely continuous foliations. If $\mathcal{F}$

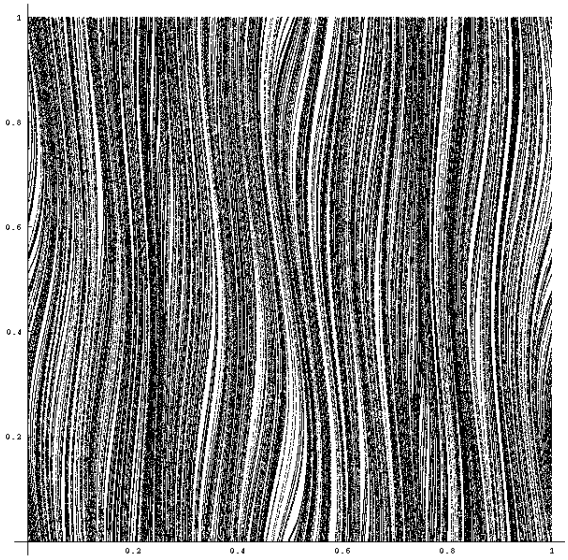


Figure 3. A pathological foliation

is a C^1 foliation, then it is absolutely continuous, but absolute continuity is a strictly weaker property.

Absolute continuity has long played a central role in smooth ergodic theory. Anosov and Sinai [1, 2] proved in the 60's that the stable and unstable foliations of globally hyperbolic (or Anosov) systems are absolutely continuous, even though they fail to be C^1 in general. Absolute continuity was a key ingredient in Anosov's celebrated proof [1] that the geodesic flow for any compact, negatively curved manifold is ergodic. When the center foliation for f fails to be absolutely continuous, this means that one cannot "quotient out by the center direction" to study ergodic properties f.

The existence of such pathological center foliations was first demonstrated by A. Katok (whose construction was written up by Milnor in [53]). Theorem E shows that this type of pathology can occur in open sets of diffeomorphisms and so is inescapable in general. In the next section, we discuss the extent to which this pathology is the norm.

Further remarks.

1. An unpublished letter of Mañé to Shub examines the consequences of nonvanishing Lyapunov center exponents on the disintegration of volume along center foliations. Some of the ideas there are pursued in greater depth in [42].

2. The examples of Katok in [53] in fact have center exponents almost everywhere equal to 0, showing that nonvanishing center exponents is not a necessary condition for atomic disintegration of volume.

3. Systems for which the center leaves are not compact (or even dynamically compact) also exhibit non-absolutely continuous center foliations, but the disintegration appears to be potentially much more complicated than just atomic disintegration [69, 36].

7. Rigidity

Examining in greater depth the potential pathologies of center foliations, we discover a rigidity phenomenon. To be concrete, let us consider the case of a perturbation $f \in \mathrm{Diff}^\infty_m(M)$ of the discretized geodesic flow on a negatively-curved surface. If the perturbation f happens to be the time-one map of a smooth flow, then $\mathcal{W}^c$ is the orbit foliation for that flow. In this case the center foliation for f is absolutely continuous – in fact, C^∞. In general, however, a perturbation f of φ_{t_0} has no reason to embed in a smooth flow, and one can ask how the volume m disintegrates along the leaves of $\mathcal{W}^c$.

There is a complete answer to this question:

Theorem F (Avila-Viana-Wilkinson [6]). *Let S be a closed negatively curved surface, and let $\varphi_t : T^1S \to T^1S$ be the geodesic flow.*

For each $t_0 > 0$, there is a neighborhood $\mathcal{U}$ of φ_{t_0} in $\mathrm{Diff}^\infty_m(T^1S)$ such that for each $f \in \mathcal{U}$:

1. *either m has atomic disintegration along the center foliation $\mathcal{W}^c$, or*
2. *f is the time-one map of a C^∞, m-preserving flow.*

What Theorem F says is that, in this context, nothing lies between C^∞ and absolute singularity of $\mathcal{W}^c$ – *pathology is all that can happen.* The geometric measure-theoretic properties of $\mathcal{W}^c$ determine completely a key dynamical property of f – whether it embeds in a flow.

The heart of the proof of Theorem F is to understand what happens when the center Lyapunov exponents *vanish*. For this, we use tools that originate in the study of random matrix products. The general theme of this work, summarized by Ledrappier in [49] is that "entropy is smaller than exponents, and entropy zero implies deterministic." Original results concerning the Lyapunov exponents of random matrix products, due to Furstenberg, Kesten [34, 33], Ledrappier [49], and others, have been extended in the past decade to deterministic products of linear cocycles over hyperbolic systems by Bonatti, Gomez-Mont, Viana [11, 13, 72]. The Bernoulli and Markov measures associated with random products in those earlier works are replaced in the newer results by invariant measures for the hyperbolic system carrying a suitable product structure.

Recent work of Avila, Viana [5] extends this hyperbolic theory from linear to *diffeomorphism* cocycles, and these results are used in a central way. For cocycles over volume preserving partially hyperbolic systems, Avila, Santamaria,

and Viana [4] have also recently produced related results, for both linear and diffeomorphism cocycles, which also play an important role in the proof. The proof in [4] employs julienne based techniques, generalizing the arguments in [24].

Further remarks.

1. The only properties of φ_{t_0} that are used in the proof of Theorem F are accessibility, dynamical coherence, one-dimensionality of E^c, the fact that φ_{t_0} fixes the leaves of $\mathcal{W}^c$, and 3-dimensionality of M. There are also more general formulations of Theorem F in [6] that relax these hypotheses in various directions. For example, a similar result holds for systems in dimension 3 for whom all center manifolds are compact.

2. Deep connections between Lyapunov exponents and geometric properties of invariant measures have long been understood [48, 50, 51, 46, 8]. Theorem F establishes new connections in the partially hyperbolic context.

3. Theorem F gives conditions under which one can recover the action of a Lie group (in this case $\mathbb{R}$) from that of a discrete subgroup (in this case $\mathbb{Z}$). These themes have arisen in the related context of measure-rigidity for algebraic partially hyperbolic actions by Einsiedler, Katok, Lindenstrauss [32]. It would be interesting to understand more deeply the connections between these works.

8. Summary, Questions

We leave this tale open-ended, with a few questions that have arisen naturally in its course.

New criteria for ergodicity. Conjecture 1 remains open. As discussed in Section 4, the julienne based techniques using the Hopf argument might have reached their limits in this problem (at least this is the case in the absence of a significantly new idea). One alternate approach which seems promising employs Lyapunov exponents and blenders [65]. Perhaps a new approach will find a satisfying conclusion to this part of the story.

Classification problem. A basic question is to understand which manifolds support partially hyperbolic diffeomorphisms. As the problem remains open in the classical Anosov case (in which E^c is zero-dimensional), it is surely exremely difficult in general. There has been significant progress in dimension 3, however; for example, using techniques in the theory of codimension-1 foliations, Burago and Ivanov proved that there are no partially hyperbolic diffeomorphisms of the 3-sphere [17].

Modifying this question slightly, one can ask whether the partially hyperbolic diffeomorphisms in low dimension must belong to certain "classes" (up to

homotopy, for example) – such as time-t maps of flows, skew products, algebraic systems, and so on. Pujals has proposed such a program in dimension 3, which has spurred several papers on the subject [15, 14, 64, 40].

It is possible that if one adds the hypotheses of dynamical coherence and absolute continuity of the center foliation, then there is such a classification. Evidence in this direction can be found in [6].

Nonuniform and singular partial hyperbolicity. Unless all of its Lyapunov exponents vanish almost everywhere, *any* volume-preserving diffeomorphism is in some sense "nonuniformly partially hyperbolic." Clearly such a general class of systems will not yield to a single approach. Nonetheless, the techniques developed recently are quite powerful and should shed some light on certain systems that are close to being partially hyperbolic. Some extensions beyond the uniform setting have been explored in [3], in which the center bunching hypotheses in [24] has been replaced by a pointwise, nonuniform center bunching condition. This gives new classes of stably ergodic diffeomorphisms that are not center bunched.

It is conceivable that the methods in [3] may be further extended to apply in certain "singular partially hyperbolic" contexts where partial hyperbolicity holds on an open, noncompact subset of the manifold M but decays in strength near the boundary. Such conditions hold, for example, for geodesic flows on certain nonpositively curved manifolds. Under suitable accessibility hypotheses, these systems should be ergodic with respect to volume.

Rigidity of partially hyperbolic actions. The rigidity phenomenon described in Section 7 has only begun to be understood. To phrase those results in a more general context, we consider a smooth, nonsingular action of an abelian Lie group G on a manifold M. Let H be a discrete group acting on M, commuting with the action of G, and whose elements are partially hyperbolic diffeomorphisms in $\mathrm{Diff}^\infty_m(M)$. Can such an action be perturbed, preserving the absolute continuity of the center foliation? How about the elements of the action? When absolute continuity fails, what happens?

The role of accessibility and accessibility classes has been exploited in a serious way in the important work of Damjanović and A. Katok on rigidity of abelian actions on quotients of $\mathrm{SL}(n,\mathbb{R})$ [26]. It seems reasonable that these explorations can be pushed further, using some of the techniques mentioned here, to prove rigidity results for other partially hyperbolic actions. A simple case currently beyond the reach of existing methods is to understand perturbations of the action of a $\mathbb{Z}^2$ lattice in the diagonal subgroup on $\mathrm{SL}(2,\mathbb{R}) \times \mathrm{SL}(2,\mathbb{R})/\Gamma$, where Γ is an irreducible lattice.

Our final question takes us further afield, but back once again to the geodesic flow. Fix a closed hyperbolic surface S, and consider the standard action on T^1S by the upper triangular subgroup $T < \mathrm{PSL}(2,\mathbb{R})$, which contains both the geodesic and positive horocyclic flows. Ghys proved that this action is highly rigid and admits no m-preserving C^∞ deformations [35]. Does the same hold true for some countable subgroup of T? For example, consider the solvable

Baumslag Solitar subgroup $BS(1,2)$ generated by the elements

$$a = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

which has the presentation $BS(1,2) = \langle a, b \mid aba^{-1} = b^2 \rangle$. Can the standard action be perturbed inside of $\mathrm{Diff}^\infty_m(T^1S)$? More generally, can one classify all faithful representations

$$\rho\colon BS(1,2) \to \mathrm{Diff}^\infty_m(M),$$

where M is a 3-manifold? For results of a similar nature in lower dimensions, see [25, 56].

References

[1] D. V. Anosov, *Geodesic flows on closed Riemannian manifolds of negative curvature*, Proc. Steklov Math. Inst. **90** (1967), 1–235.

[2] D. V. Anosov and Ya. G. Sinai, *Certain smooth ergodic systems*, Russian Math. Surveys **22** (1967), 103–167.

[3] A. Avila, J. Bochi, and A. Wilkinson, *Nonuniform center bunching and the genericity of ergodicity among C^1 partially hyperbolic symplectomorphisms.*

[4] A. Avila, J. Santamaria, and M. Viana, *Cocycles over partially hyperbolic maps*, Preprint www.preprint.impa.br 2008.

[5] A. Avila and M. Viana, *Extremal Lyapunov exponents of smooth cocycles*, Invent. Math.

[6] A. Avila, M. Viana, and A. Wilkinson, *Absolute continuity, Lyapunov exponents, and rigidity*, in preparation.

[7] A. Baraviera and C. Bonatti, *Removing zero central Lyapunov exponents*, Ergod. Th. & Dynam. Sys. **23** (2003), 1655–1670.

[8] L. Barreira, Ya. Pesin, and J. Schmeling, *Dimension and product structure of hyperbolic measures*, Ann. of Math. **149** (1999), 755–783.

[9] C. Bonatti and L. J. Díaz, *Nonhyperbolic transitive diffeomorphisms*, Ann. of Math. **143** (1996), 357–396.

[10] C. Bonatti, L. J. Díaz, and M. Viana, *Dynamics beyond uniform hyperbolicity*, Encyclopaedia of Mathematical Sciences, vol. 102, Springer-Verlag, 2005.

[11] C. Bonatti, X. Gómez-Mont, and M. Viana, *Généricité d'exposants de Lyapunov non-nuls pour des produits déterministes de matrices*, Ann. Inst. H. Poincaré Anal. Non Linéaire **20** (2003), 579–624.

[12] C. Bonatti, C. Matheus, M. Viana, and A. Wilkinson, *Abundance of stable ergodicity*, Comment. Math. Helv. **79** (2004), 753–757.

[13] C. Bonatti and M. Viana, *Lyapunov exponents with multiplicity 1 for deterministic products of matrices*, Ergod. Th. & Dynam. Sys **24** (2004), 1295–1330.

[14] C. Bonatti and A. Wilkinson, *Transitive partially hyperbolic diffeomorphisms on 3-manifolds*, Topology **44** (2005), 475–508.

[15] M. Brin, D. Burago, and S. Ivanov, *On partially hyperbolic diffeomorphisms on* 3*-manifolds with commutative fundamental group*, Advances in Dynamical Systems, Cambridge Univ. Press.

[16] M. Brin and Ya. Pesin, *Partially hyperbolic dynamical systems*, Izv. Acad. Nauk. SSSR **1** (1974), 177–212.

[17] D. Burago and S. Ivanov, *Partially hyperbolic diffeomorphisms of 3-manifolds with abelian fundamental groups*, J. Mod. Dyn. **2** (2008), 541–580.

[18] K. Burns, D. Dolgopyat, and Ya. Pesin, *Partial hyperbolicity, Lyapunov exponents and stable ergodicity*, J. Statist. Phys. **108** (2002), 927–942, Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays.

[19] K. Burns, F. Rodríguez Hertz, M. A. Rodríguez Hertz, A. Talitskaya, and R. Ures, *Density of accessibility for partially hyperbolic diffeomorphisms with one-dimensional center*, Discrete Contin. Dyn. Syst. **22** (2008), 75–88.

[20] K. Burns, C. Pugh, M. Shub, and A. Wilkinson, *Recent results about stable ergodicity*, Smooth ergodic theory and its applications (Seattle WA, 1999), Procs. Symp. Pure Math., vol. 69, Amer. Math. Soc., 2001, pp. 327–366.

[21] K. Burns, C. Pugh, and A. Wilkinson, *Stable ergodicity and Anosov flows*, Topology **39** (2000), 149–159.

[22] K. Burns and A. Wilkinson, *On the ergodicity of partially hyperbolic systems*, Ann. of Math.

[23] ———, *Stable ergodicity of skew products*, Ann. Sci. École Norm. Sup. **32** (1999), 859–889.

[24] ———, *Dynamical coherence and center bunching*, Discrete Contin. Dyn. Syst. **22** (2008), 89–100.

[25] L. Burslem and A. Wilkinson, *Global rigidity of solvable group actions on* S^1, Geometry and Topology **8** (2004), 877–924.

[26] D. Damjanović and A. Katok, *Local rigidity of partially hyperbolic actions. II. The geometric method and restrictions of weyl chamber flows on* $\mathrm{SL}(n,\mathbb{R})/L$.

[27] L. J. Díaz, E. Pujals, and R. Ures, *Partial hyperbolicity and robust transitivity*, Acta Math. **183** (1999), 1–43.

[28] P. Didier, *Stability of accessibility*, Ergod. Th. & Dynam. Sys. **23** (2003), 1717–1731.

[29] D. Dolgopyat, *On decay of correlations in Anosov flows*, Ann. of Math. **147** (1998), 357–390.

[30] ———, *On differentiability of SRB states for partially hyperbolic systems*, Invent. Math. **155** (2004), 389–449.

[31] D. Dolgopyat and A. Wilkinson, *Stable accessibility is* C^1 *dense*, Astérisque **287** (2003), 33–60.

[32] M. Einsiedler, A. Katok, and E. Lindenstrauss, *Invariant measures and the set of exceptions to Littlewood's conjecture*, Ann. of Math. **164** (2006), 513–560.

[33] H. Furstenberg, *Non-commuting random products*, Trans. Amer. Math. Soc. **108** (1963), 377–428.

[34] H. Furstenberg and H. Kesten, *Products of random matrices*, Ann. Math. Statist. **31** (1960), 457–469.

[35] É. Ghys, *Actions localement libres du groupe affine*, Invent. Math. **82** (1985), 479–526.

[36] A. Gogolev, *How typical are pathological foliations in partially hyperbolic dynamics: an example*, arXiv:0907.3533.

[37] N. Gourmelon, *Adapted metrics for dominated splittings*, Ergod. Th. & Dynam. Sys. **27** (2007), 1839–1849.

[38] M. Grayson, C. Pugh, and M. Shub, *Stably ergodic diffeomorphisms*, Ann. of Math. **140** (1994), 295–329.

[39] J. Hadamard, *Les surfaces à courbures opposées et leurs lignes géodésiques*, Journal de Math. Pures et Appliquées **IV** (1898), 27–73.

[40] A. Hammerlindl, *Leaf conjugacies on the torus*, Ph.D. thesis, U. Toronto, 2009.

[41] B. Hasselblatt and Ya. Pesin, *Partially hyperbolic dynamical systems*, Handbook of dynamical systems. Vol. 1B, Elsevier B. V., 2006, pp. 1–55.

[42] M. Hirayama and Ya. Pesin, *Non-absolutely continuous foliations*, Israel J. Math. **160** (2007), 173–187.

[43] M. Hirsch, C. Pugh, and M. Shub, *Invariant manifolds*, Lect. Notes in Math., vol. 583, Springer Verlag, 1977.

[44] E. Hopf, *Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung*, Ber. Verh. Sächs. Akad. Wiss. Leipzig **91** (1939), 261–304.

[45] V. Horita and A. Tahzibi, *Partial hyperbolicity for symplectic diffeomorphisms*, Ann. Inst. H. Poincaré Anal. Non Linéaire **23** (2006), 641–661.

[46] A. Katok, *Lyapunov exponents, entropy and periodic points of diffeomorphisms*, Publ. Math. IHES **51** (1980), 137–173.

[47] A. Katok and A. Kononenko, *Cocycle's stability for partially hyperbolic systems*, Math. Res. Lett. **3** (1996), 191–210.

[48] F. Ledrappier, *Propriétés ergodiques des mesures de Sinaï*, Publ. Math. I.H.E.S. **59** (1984), 163–188.

[49] ———, *Positivity of the exponent for stationary sequences of matrices*, Lyapunov exponents (Bremen, 1984), Lect. Notes Math., vol. 1186, Springer, 1986, pp. 56–73.

[50] F. Ledrappier and L.-S. Young, *The metric entropy of diffeomorphisms. I. Characterization of measures satisfying Pesin's entropy formula*, Ann. of Math. **122** (1985), 509–539.

[51] ———, *The metric entropy of diffeomorphisms. II. Relations between entropy, exponents and dimension*, Ann. of Math. **122** (1985), 540–574.

[52] C. Liverani, *On contact Anosov flows*, Ann. of Math. **159** (2004), 1275–1312.

[53] J. Milnor, *Fubini foiled: Katok's paradoxical example in measure theory*, Math. Intelligencer **19** (1997), 30–32.

[54] V. Niţică and A. Török, *An open dense set of stably ergodic diffeomorphisms in a neighborhood of a non-ergodic one*, Topology **40** (2001), 259–278.

[55] Ya. Pesin, *Lectures on partial hyperbolicity and stable ergodicity*, European Mathematical Society (EMS), 2006, Zurich Lectures in Advanced Mathematics.

[56] L. Polterovich, *Growth of maps, distortion in groups and symplectic geometry*, Invent. Math. **150** (2002), 655686.

[57] C. Pugh and M. Shub, *Ergodicity of Anosov actions*, Invent. Math. **15** (1972), 1–23.

[58] ———, *Stable ergodicity and partial hyperbolicity*, International Conference on Dynamical Systems (Montevideo, 1995), Pitman Res. Notes Math. Ser., vol. 362, Longman, 1996, pp. 182–187.

[59] ———, *Stably ergodic dynamical systems and partial hyperbolicity*, J. Complexity **13** (1997), 125–179.

[60] ———, *Stable ergodicity and julienne quasi-conformality*, J. Europ. Math. Soc. **2** (2000), 1–52.

[61] F. Rodríguez Hertz, *Stable ergodicity of certain linear automorphisms of the torus*, Ann. of Math. **162** (2005), 65–107.

[62] F. Rodríguez Hertz, M. A. Rodríguez Hertz, and R. Ures, *A survey of partially hyperbolic dynamics*, Partially hyperbolic dynamics, laminations, and Teichmüller flow, Fields Inst. Commun., vol. 51, Amer. Math. Soc., 2007, pp. 35–87.

[63] ———, *Accessibility and stable ergodicity for partially hyperbolic diffeomorphisms with 1D-center bundle*, Invent. Math. **172** (2008), 353–381.

[64] ———, *Partial hyperbolicity and ergodicity in dimension three*, J. Mod. Dyn. **2** (2008), 187–208.

[65] J. Rodríguez Hertz, M. A. Rodríguez Hertz, A. Tahzibi, and R. Ures, *A criterion for ergodicity of non-uniformly hyperbolic diffeomorphisms.*, www.arXiv.org arXiv:0710.2353.

[66] D. Ruelle, *Perturbation theory for Lyapunov exponents of a toral map: extension of a result of Shub and Wilkinson*, Israel J. Math. **134** (2003), 345–361.

[67] D. Ruelle and A. Wilkinson, *Absolutely singular dynamical foliations*, Comm. Math. Phys. **219** (2001), 481–487.

[68] R. Saghin and Z. Xia, *Partial hyperbolicity or dense elliptic periodic points for c^1-generic symplectic diffeomorphisms*, Trans. Amer. Math. Soc. **358** (2006), 5119–5138.

[69] ———, *Geometric expansion, Lyapunov exponents and foliations*, Ann. Inst. H. Poincaré Anal. Non Linéaire **26** (2009), 689–704.

[70] M. Shub and A. Wilkinson, *Pathological foliations and removable zero exponents*, Invent. Math. **139** (2000), 495–508.

[71] A. Tahzibi, *Stably ergodic diffeomorphisms which are not partially hyperbolic*, Israel J. Math. **142** (2004), 315–344.

[72] M. Viana, *Almost all cocycles over any hyperbolic system have nonvanishing Lyapunov exponents*, Ann. of Math. **167** (2008), 643–680.

[73] A. Wilkinson, *The cohomological equation for partially hyperbolic diffeomorphisms*, arXiv:0809.4862.

[74] ———, *Stable ergodicity of the time-one map of a geodesic flow*, Ergod. Th. & Dynam. Sys. **18** (1998), 1545–1587.

Section 11

Partial Differential Equations

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

A Hyperbolic Dispersion Estimate, with Applications to the Linear Schrödinger Equation

Nalini Anantharaman*

Abstract

On a Hilbert space $\mathcal{H}$, consider the product $\hat{P}_n\hat{P}_{n-1}\cdots\hat{P}_1$ of a large number of operators $\hat{P}_j$, with $\|\hat{P}_j\| = 1$. What kind of geometric considerations can serve to prove that the norm $\|\hat{P}_n\hat{P}_{n-1}\cdots\hat{P}_1\|$ decays exponentially fast with n ? In the first part of this note, we will describe a situation in which $\mathcal{H} = L^2(\mathbb{R}^d)$, and the operators $\hat{P}_j$ are Fourier integral operators associated to a sequence of canonical transformations κ_j. We will give conditions, on the sequence of transformations κ_j and on the symbols of the operators $\hat{P}_j$, under which we can prove exponential decay. This technique was introduced to prove results related to the quantum unique ergodicity conjecture. In the second half of this paper, we will survey applications in scattering situations, to prove the existence of a gap below the real axis in the resolvent spectrum, and to get local smoothing estimates with loss, as well as Strichartz estimates.

Mathematics Subject Classification (2010). Primary 35P20; Secondary 37D99.

Keywords. Quantum chaos, Schrödinger equation, quantum unique ergodicity, hyperbolic dynamical systems, resonances, Strichartz estimates

1. Introduction

On a Hilbert space $\mathcal{H}$, consider the product $\hat{P}_n\hat{P}_{n-1}\cdots\hat{P}_1$ of a large number of operators $\hat{P}_j$, with $\|\hat{P}_j\| = 1$. Think, for instance, of the case where each operator $\hat{P}_j$ is an orthogonal projector, or a product of an orthogonal projector

*The author wishes to acknowledge the support of Agence Nationale de la Recherche, under the grant ANR-05-JCJC-0107-01 followed by ANR-09-JCJC-0099-01. She also thanks Peter Sarnak and Yves Colin de Verdière for their continued support, and Lior Silberman and Stéphane Nonnenmacher with whom it has been a pleasure to work.

Département de Mathématiques, Bâtiment 425, Faculté des Sciences d'Orsay, Université Paris-Sud, F-91405 Orsay Cedex. E-mail: Nalini.Anantharaman@math.u-psud.fr.

and a unitary operator. What kind of geometric considerations can be helpful to prove that the norm $\|\hat{P}_n\hat{P}_{n-1}\cdots\hat{P}_1\|$ is strictly less than 1 ? or better, that it decays exponentially fast with n ? In Section 2, we will describe a situation in which $\mathcal{H} = L^2(\mathbb{R}^d)$, and the operators $\hat{P}_j$ are Fourier integral operators associated to a sequence of canonical transformations κ_j. We will give a "hyperbolicity" condition, on the sequence of transformations κ_j and on the symbols of the operators $\hat{P}_j$, under which we can prove exponential decay of the norm $\|\hat{P}_n\hat{P}_{n-1}\cdots\hat{P}_1\|$. This technique was introduced in [1, 2], and was used in [1, 2, 3, 29, 4] to prove results related to the quantum unique ergodicity conjecture, for eigenfunctions of the laplacian on negatively curved manifolds: see Section 3. In the last section of this paper (Section 4), we will survey the work of Nonnenmacher-Zworski [27, 28], Christianson [8, 9, 10], Datchev [12], and Burq-Guillarmou-Hassell [6], who showed how to use the previous estimates in scattering situations, to prove the existence of a gap below the real axis in the resolvent spectrum, and to get local smoothing estimates with loss, as well as Strichartz estimates.

2. The Hyperbolic Dispersion Estimate

In this section, $\mathbb{R}^d \times (\mathbb{R}^d)^*$ is endowed with the canonical symplectic form $\omega = \sum_{j=1}^d dx_j \wedge d\xi_j$, where dx_j denotes the projection on the j-th vector of the canonical basis in $\mathbb{R}^d$, and $d\xi_j$ is the projection on the j-th vector of the dual basis in $(\mathbb{R}^d)^*$. The space $\mathbb{R}^d$ will also be endowed with its usual scalar product, denoted $\langle .,. \rangle$, and we will use it to systematically identify $\mathbb{R}^d$ with $(\mathbb{R}^d)^*$.

We consider a sequence of smooth ($\mathcal{C}^\infty$) canonical transformations $\kappa_n : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^d \times \mathbb{R}^d$, preserving ω. We will only be interested in the restriction of κ_1 to a fixed relatively compact neighbourhood Ω of 0, and it is actually sufficient for us to assume that the product $\kappa_n \circ \kappa_{n-1} \circ \cdots \circ \kappa_1$ is well defined, for all n, on Ω. The Darboux-Lie theorem ensures that every lagrangian foliation can be mapped, by a symplectic change of coordinates, to the foliation of $\mathbb{R}^d \times \mathbb{R}^d$ by the "horizontal" leaves $\mathcal{L}_{\xi_0} = \{(x,\xi) \in \mathbb{R}^d \times \mathbb{R}^d, \xi = \xi_0\}$. Thus, for our purposes, there is no loss of generality if we make the simplifying assumption that each symplectic transformation κ_n preserves this horizontal foliation. It means that κ_n is of the form $(x,\xi) \mapsto (x', \xi' = p_n(\xi))$ where $p_n : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is a smooth function. In more elaborate words, κ_n has a generating function of the form $S_n(x,x',\theta) = \langle p_n(\theta), x' \rangle - \langle \theta, x \rangle + \alpha_n(\theta)$ (where $x, x', \theta \in \mathbb{R}^d$, and $\alpha_n : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is a smooth function). We have the equivalence

$$\begin{aligned}\big[(x',\xi') \quad = \quad \kappa_n(x,\xi)\big] \Longleftrightarrow \big[\xi = -\partial_x S_n(x,x',\theta),\ \xi' \\ = \partial_{x'} S_n(x,x',\theta),\ \partial_\theta S_n(x,x',\theta) = 0\big].\end{aligned}$$

The product $\kappa_n \circ \ldots \circ \kappa_2 \circ \kappa_1$ also preserves the horizontal foliation, and it

admits the generating function

$$\langle p_n \circ \ldots \circ p_1(\theta), x' \rangle - \langle \theta, x \rangle + \alpha_1(\theta) + \alpha_2(p_1(\theta)) + \ldots + \alpha_n(p_{n-1} \circ \ldots \circ p_1(\theta)) \\ = \langle p_n \circ \ldots \circ p_1(\theta), x' \rangle - \langle \theta, x \rangle + A_n(\theta),$$

where the equality defines $A_n(\theta)$.

We will assume that the functions p_n are smooth diffeomorphisms, and that all the derivatives of p_n, of p_n^{-1} and of α_n are bounded uniformly in n. If p is a map $\mathbb{R}^d \longrightarrow \mathbb{R}^d$, we will denote ∇p the matrix $(\frac{\partial p_i}{\partial \theta_j})_{ij}$, which represents its differential in the canonical basis.

Assumptions (H): We shall be interested in the following operators, acting on $L^2(\mathbb{R}^d)$:

$$\hat{P}_n f(x') = \frac{1}{(2\pi\hbar)^d} \int_{x \in \mathbb{R}^d, \theta \in \mathbb{R}^d} e^{\frac{iS_n(x,x',\theta)}{\hbar}} a^{(n)}(x, x', \theta, \hbar) f(x) dx d\theta,$$

where $\hbar > 0$ is a parameter destined to go to 0. We will assume that the functions $a^{(n)}(x, x', \theta, \hbar)$ have the following properties:

- For a given $\hbar > 0$, the function $(x, x', \theta) \mapsto a^{(n)}(x, x', \theta, \hbar)$ is of class $\mathcal{C}^\infty$;
- The function $a^{(1)}(x, x', \theta, \hbar)$ is supported in Ω with respect to the variable x;
- With respect to the variables (x', θ), the functions $a^{(n)}(x, x', \theta, \hbar)$ have a compact support $x' \in \Omega_1, \theta \in \Omega_2$, independent of n and $\hbar$;
- When $\hbar \longrightarrow 0$, each $a^{(n)}(x, x', \theta, \hbar)$ has an asymptotic expansion

$$a^{(n)}(x, x', \theta, \hbar) \sim (\det \nabla p_n(\theta))^{1/2} \sum_{k=0}^{\infty} \hbar^k a_k^{(n)}(x, x', \theta),$$

  valid up to any order and in all the $\mathcal{C}^\ell$ norms. Besides, these asymptotic expansions are uniform with respect to n.

- If $(x', \theta') = \kappa_n(x, \theta)$, we have $|a_0^{(n)}(x, x', \theta)| \leq 1$. This condition ensures that $\|\hat{P}_n\|_{L^2 \longrightarrow L^2} \leq 1 + \mathcal{O}(\hbar)$.

The operators $\hat{P}_n$ are (semiclassical) *Fourier integral operators* associated with the transformations κ_n.

2.1. Propagation of a single plane wave.

The following theorem is essentially proved in [1]. We denote $e_{\xi_0,\hbar}$ the function $e_{\xi_0,\hbar}(x) = e^{\frac{i\langle \xi_0, x \rangle}{\hbar}}$.

Theorem 2.1. *Fix $\xi_0 \in \mathbb{R}^d$. In addition to the assumptions above, assume that*

$$\limsup_{n\longrightarrow+\infty} \frac{1}{n} \log\|\nabla(p_n \circ \ldots \circ p_2 \circ p_1)(\xi_0)\| \leq 0.$$

Fix $\mathcal{K} > 0$ arbitrary, and an integer $M \in \mathbb{N}$. Then we have, for $n = \mathcal{K}|\log\hbar|$,

$$\hat{P}_n \circ \ldots \circ \hat{P}_2 \circ \hat{P}_1 e_{\xi_0,\hbar}(x)$$
$$= e^{i\frac{A_n(\xi_0)}{\hbar}} e_{\xi_n,\hbar}(x)(\det \nabla p_n \circ \ldots \circ p_1(\xi_0))^{1/2} \left[\sum_{k=0}^{M-1} \hbar^k b_k^{(n)}(x,\xi_n)\right]$$
$$+\mathcal{O}(\hbar^M).$$

The functions $b_k^{(n)}$, defined on $\mathbb{R}^d \times \mathbb{R}^d$, are smooth, and

$$b_0^{(n)}(x_n,\xi_n) = \prod_{j=1}^{n} a_0^{(j)}(x_j, x_{j+1}, \xi_j),$$

where we denote $\xi_n = p_n \circ \ldots \circ p_1(\xi_0)$, $x_n = x$ and the other terms are defined by the relations $(x_j,\xi_j) = \kappa_j \circ \ldots \circ \kappa_1(x_0,\xi_0)$.

The next terms $b_k^{(n)}$ have the same support as $b_0^{(n)}$. We have $|b_0^{(n)}(x_n,\xi_n)| \leq 1$, and besides, we have bounds

$$\|d^m b_k^{(n)}\| \leq C(k,m) n^{m+3k},$$

where $C(k,m)$ does not depend on n.

If n is fixed, and if we write $\hat{P}_n \circ \ldots \circ \hat{P}_2 \circ \hat{P}_1 e_{\xi_0,\hbar}(x)$ explicitly as an integral over $(\mathbb{R}^d)^{2n}$, this theorem is a straightforward application of the stationary phase method. If n is allowed to go to infinity as $\hbar \longrightarrow 0$, our result amounts to applying the method of stationary phase on a space whose dimension goes to ∞, and this is known to be very delicate. The theorem was first proved this way, in an unpublished version (available on request or on my webpage) of the paper [1]. A nicer proof is available in [1], and has also appeared under different forms in [2, 27]. In these papers, the proofs are written on a riemannian manifold, for $\hat{P}_n = e^{\frac{i\tau\hbar\Delta}{2}}\hat{\chi}_n$, where the operators $\hat{\chi}_n$ belong to a finite family of pseudodifferential operators, whose symbols are supported inside compact sets of small diameters, and where Δ is the laplacian and $\tau > 0$ is fixed. In local coordinates, the calculations done in [1, 2, 27] amount to the simpler statement presented here. In the unpublished version, the assumptions were much stronger; the transformations κ_j were assumed to be analytic, and the symbols $a^{(n)}$ were taken in a Gevrey class. The result was also much stronger, in that the conclusion held for $n = \hbar^{-\delta}$, for some $\delta > 0$.

In all the papers under review, the dynamical systems under study satisfy a uniform hyperbolicity condition, ensuring an exponential decay

$$\sup_{\xi\in\Omega_2} \|\nabla(p_n \circ \ldots \circ p_2 \circ p_1)(\xi)\| \leq Ce^{-\lambda n}, \tag{1}$$

with fixed constants $C, \lambda > 0$. This is why, following [27], we call our result a *hyperbolic dispersion estimate*. Applications will be surveyed in Sections 3 and 4.

2.2. Estimating the norm of $\hat{P}_n \circ \ldots \circ \hat{P}_2 \circ \hat{P}_1$. We use the $\hbar$-Fourier transform

$$\mathcal{F}_\hbar u(\xi) = \frac{1}{(2\pi\hbar)^{d/2}} \int_{\mathbb{R}^d} u(x) e^{-\frac{i\langle \xi, x\rangle}{\hbar}} dx,$$

the inversion formula

$$u(x) = \frac{1}{(2\pi\hbar)^{d/2}} \int_{\mathbb{R}^d} \mathcal{F}_\hbar u(\xi) e^{\frac{i\langle \xi, x\rangle}{\hbar}} d\xi,$$

and the Plancherel formula $\|u\|_{L^2(\mathbb{R}^d)} = \|\mathcal{F}_\hbar u\|_{L^2(\mathbb{R}^d)}$. Using the Fourier inversion formula, Theorem 2.1 implies, in a straightforward manner, the following

Theorem 2.2. *In addition to the assumptions above, assume that*

$$\limsup_{n \longrightarrow +\infty} \frac{1}{n} \log \|\nabla (p_n \circ \ldots \circ p_2 \circ p_1)(\xi)\| \leq 0,$$

uniformly in $\xi \in \Omega_2$.

Fix $\mathcal{K} > 0$ arbitrary. Then there exists $\hbar_\mathcal{K} > 0$ such that, for $n = \mathcal{K} |\log \hbar|$, and for $\hbar < \hbar_\mathcal{K}$,

$$\|\hat{P}_n \circ \ldots \circ \hat{P}_2 \circ \hat{P}_1\|_{L^2 \longrightarrow L^2} \leq \frac{|\Omega_2|^{1/2}}{(2\pi\hbar)^{d/2}} \sup_{\xi \in \Omega_2} |\det \nabla p_n \circ \ldots \circ p_1(\xi)|^{1/2} (1 + \mathcal{O}(n^3 \hbar)),$$

where $|\Omega_2|$ denotes the volume of Ω_2.

Of course, we always have the trivial bound $\|\hat{P}_n \circ \ldots \circ \hat{P}_2 \circ \hat{P}_1\|_{L^2 \longrightarrow L^2} \leq 1 + \mathcal{O}(\hbar |\log \hbar|)$. Since we are working in the limit where $\hbar \longrightarrow 0$, our estimate can only have an interest if we have an upper bound of the form

$$\sup_{\xi \in \Omega_2} |\det \nabla p_n \circ \ldots \circ p_1(\xi)|^{1/2} \leq C e^{-\lambda n}, \qquad \lambda > 0, \tag{2}$$

and if $\mathcal{K}$ is large enough. Note that (2) is weaker than the condition (1).

We now state a refinement of Theorem 2.2. We consider the same family $\hat{P}_i$, satisfying Assumptions (H). The multiplicative constants in our estimate have no importance, and in what follows we will omit them.

Theorem 2.3. *[4] In addition to the assumptions above, assume that*

$$\limsup_{n \longrightarrow +\infty} \frac{1}{n} \log \|\nabla (p_n \circ \ldots \circ p_2 \circ p_1)(\xi)\| \leq 0,$$

uniformly in $\xi \in \Omega_2$.

Let $r \leq d$, and assume that the coisotropic foliation by the leaves $\{\xi_{r+1} = c_{r+1}, \dots, \xi_d = c_d\}$ is invariant by each canonical transformation κ_n. In other words, the map p_n is of the form

$$p_n((\xi_1, \dots, \xi_r), (\xi_{r+1}, \dots, \xi_d)) = (m_n(\xi_1, \dots, \xi_d), \tilde{p}_n(\xi_{r+1}, \dots, \xi_d)),$$

where $m_n : \mathbb{R}^d \longrightarrow \mathbb{R}^r$ and $\tilde{p}_n : \mathbb{R}^{d-r} \longrightarrow \mathbb{R}^{d-r}$.

Fix $\mathcal{K} > 0$ arbitrary. Then there exists $\hbar_{\mathcal{K}} > 0$ such that, for $n = \mathcal{K}|\log \hbar|$, and for $\hbar < \hbar_{\mathcal{K}}$,

$$\|\hat{P}_n \circ \dots \circ \hat{P}_2 \circ \hat{P}_1\|_{L^2 \longrightarrow L^2} \leq \frac{1}{(2\pi\hbar)^{r/2}} \frac{\sup_{\xi \in \Omega_2} |(\det \nabla p_n \circ \dots \circ p_1(\xi))|^{1/2}}{\inf_{\xi \in \Omega_2} |(\det \nabla \tilde{p}_n \circ \dots \circ \tilde{p}_1(\xi))|^{1/2}} (1 + \mathcal{O}(n^3 \hbar)).$$

Theorem 2.3 is an improvement of Theorem 2.2 in the case where we have

$$\frac{1}{(2\pi\hbar)^{d/2}} \sup_{\xi \in \Omega_2} |(\det \nabla p_n \circ \dots \circ p_1(\xi_0))^{1/2}| \gg 1$$

but

$$\frac{1}{(2\pi\hbar)^{r/2}} \frac{\sup_{\xi \in \Omega_2} |(\det \nabla p_n \circ \dots \circ p_1(\xi))|^{1/2}}{\inf_{\xi \in \Omega_2} |(\det \nabla \tilde{p}_n \circ \dots \circ \tilde{p}_1(\xi))|^{1/2}} \ll 1.$$

As a trivial example, when each κ_n is the identity, Theorem 2.2 gives a non-optimal bound, whereas we can take $r = 0$ in Theorem 2.3, and recover the (almost) optimal bound $\|\hat{P}_n \circ \dots \circ \hat{P}_2 \circ \hat{P}_1\|_{L^2 \longrightarrow L^2} \leq 1 + \mathcal{O}(\hbar |\log \hbar|^3)$. A less trivial example will appear in Section 3.

3. An Application to the Quantum Unique Ergodicity Conjecture

3.1. Statement of the conjecture. Let X be a d-dimensional compact riemannian manifold, let Δ denote the Laplace-Beltrami operator on X, and let V be a smooth function on X. In the most general framework, the question of "quantum ergodicity" asks about the behaviour of the solutions of the stationary Schrödinger equation

$$\left(-\hbar^2 \frac{\Delta}{2} + V\right) \psi_\hbar = E_\hbar \psi_\hbar, \tag{3}$$

in the limit $\hbar \longrightarrow 0$ and assuming the eigenvalue $E_\hbar$ converges to a fixed value E. We will always assume that the eigenfunction $\psi_\hbar$ is normalized in $L^2(X, \mathrm{Vol})$. Quantum ergodicity asks about the weak limits of the family of probability measures $|\psi_\hbar(x)|^2 d\mathrm{Vol(x)}$. Actually, people are interested in a family of distributions $\mu_\hbar$ on the cotangent bundle T^*X, that contain more information, defined as follows:

$$\forall a \in \mathcal{C}_c^\infty(T^*X), \langle \mu_\hbar, a \rangle = \langle \psi_\hbar, \mathrm{Op}_\hbar(a) \psi_\hbar \rangle_{L^2(X)}, \tag{4}$$

where $\mathrm{Op}_\hbar(a)$ is a semiclassical pseudodifferential operator with principal symbol a (if $a = a(x,\xi)$, then $\mathrm{Op}_\hbar(a) = a(x, -i\hbar\partial_x)$, and this can be defined properly using the Weyl calculus in local coordinates). The distribution $\mu_\hbar$ appears under various names in the literature, depending on the specific context: Wigner transform, semiclassical/microlocal defect measure, microlocal lift of $\psi_\hbar$... Although the definition of $\mu_\hbar$ depends on the choice of local coordinates, the collection of weak limits of $\mu_\hbar$, as $\hbar \longrightarrow 0$, is well defined, independently on any choices. Besides, the definition (4) can be extended to the case when a is a function on T^*X depending only on the base point x, and in that case $\mathrm{Op}_\hbar(a)$ is the multiplication operator by a. We see that the projection of $\mu_\hbar$ on X is the probability measure $|\psi_\hbar(x)|^2 d\mathrm{Vol(x)}$ that we were originally interested in. The distribution $\mu_\hbar$ contains more information, it tells us something about the local directions of oscillations of $\psi_\hbar$.

The following is a form of the theorem of propagation of singularities, due to Hörmander. Define the function $H(x,\xi) = \frac{\|\xi\|_x^2}{2} + V(x)$, on T^*X – where $\|.\|_x^2$ is the norm on T_x^*X dual to the riemannian metric. Denote (Φ_H^t) the hamiltonian flow defined by H, acting on T^*X. In local coordinates, the flow (Φ_H^t) is defined by the following first order differential equation:

$$\begin{cases} \dot{x} = \frac{\partial H}{\partial \xi} \\ \\ \dot{\xi} = -\frac{\partial H}{\partial x}\,. \end{cases} \tag{5}$$

We will denote by Y_H, or simply Y, the vector field on T^*X associated with this flow.

Theorem 3.1. *(i) Given any sequence $\hbar_n \longrightarrow 0$, one can extract from the sequence $(\mu_{\hbar_n})$ a converging subsequence in $\mathcal{D}'(T^*X)$.*

We will call limits of such subsequences "semiclassical measures" associated with the family $(\psi_\hbar)$.

(ii) Let μ be a semiclassical measure. Then μ is a probability measure, carried by the level set $\{H = E\}$.

(iii) In addition, μ is invariant by the hamiltonian flow (Φ_H^t): we have $(\Phi_H^t)_*\mu = \mu$, for all t.

This theorem does not suffice to characterize a unique limit μ, as there are generally many invariant measures under (Φ_H^t). A hamiltonian flow on T^*X always preserves the Liouville measure, defined in local coordinates by $dxd\xi$: this measure, or more precisely its disintegration on $\{H = E\}$, is a candidate to be a semiclassical measure. If the flow (Φ_H^t) has periodic orbits on the energy level $\{H = E\}$, each of them carries an invariant measure, which is also a candidate to be a semiclassical measure. Characterizing the set of semiclassical measures is, in such generality, an open question. The two most studied cases are *completely integrable* hamiltonian flows on the one hand, *"chaotic"* flows on the other hand. In what follows we will focus on the "chaotic" case, and will give a more precise definition of this term.

Until the end of this section, we turn to a special case which has been most studied, and is a source of numerous open questions: the case when $V = 0$. In this case, (Φ_H^t) is the geodesic flow; we shall simply denote it by (Φ^t). We consider the case of a non-singular energy level, in other words $E \neq 0$, and since in this case the function H is homogeneous with respect to ξ, we may decide without loss of generality to take $E = \frac{1}{2}$. Then the level set $\{H = E\}$ is the unit cotangent bundle S^*X. Letting $\lambda = \frac{E_\hbar}{\hbar^2}$, equation (3) amounts to studying the eigenfunctions of the laplacian,

$$-\Delta\phi_\lambda = \lambda\phi_\lambda,$$

in the limit $\lambda \longrightarrow +\infty$. We recall that, on a compact manifold, the eigenvalues λ form a discrete set. We denote $\mu_\lambda \in \mathcal{D}'(T^*X)$ the distribution defined previously, by $\langle \mu_\lambda, a\rangle = \langle \phi_\lambda, \mathrm{Op}_{\lambda^{-1/2}}(a)\phi_\lambda\rangle$. We can rephrase our question by asking: among the invariant probability measures of the geodesic flow, which ones can be obtained as limits of the family (μ_λ) ? does the answer depend on the geometry ? The following theorem is referred to as "the Shnirelman theorem", or "the quantum ergodicity theorem". It was later extended to more general hamiltonian flows [20], and to the case of manifolds with a boundary (when X has a boundary, one has to impose boundary conditions to the eigenfunctions) [18].

Theorem 3.2. *[33, 38, 11] Let X be a compact riemannian manifold. Let (ϕ_n) be an orthonormal basis of $L^2(X)$ formed by eigenfunctions of the laplacian ($-\Delta\phi_n = \lambda_n\phi_n$, with $\lambda_n \longrightarrow +\infty$). Denote $\mu_n = \mu_{\lambda_n}$.*

Assume that the geodesic flow, acting on the unit cotangent bundle S^*X, is ergodic with respect to the Liouville measure. Then, there exists a subset $\mathcal{S} \subset \mathbb{N}$ of density 1, such that

$$\mu_n \overset{n\in\mathcal{S}}{\longrightarrow} \text{Liouville},$$

the convergence taking place in $\mathcal{D}'(T^*X)$.

The set $\mathcal{S}$ being of density 1 means that $\frac{\sharp\mathcal{S}\cap[0,N]}{N} \underset{N\to+\infty}{\longrightarrow} 1$.

It is a difficult question to know whether the whole sequence (μ_n) converges, or if there can be exceptional subsequences converging to a measure other than Liouville. Of course, the answer depends on the geometry. A particularly frustrating example is the case where X is a euclidean domain in $\mathbb{R}^2$ in the shape of a stadium, called the Bunimovich stadium. In this example, it is quite clear in numerical simulations that, although Shnirelman's theorem holds, there are also exceptional subsequences concentrating on the periodic trajectories that bounce back and forth between the two parallel sides of the stadium. The first breakthrough in that direction was made in 2008 by Hassell [19], who showed, for "almost all stadia", that there are exceptional subsequences of eigenfunctions.

If X is a compact riemannian manifold with negative sectional curvatures, Rudnick and Sarnak conjectured that, for any orthonormal basis of eigenfunctions (ϕ_n), the whole sequence (μ_n) should converge to the Liouville measure:

this is referred to as the quantum unique ergodicity conjecture [31]. A special case of this conjecture, called arithmetic quantum unique ergodicity, was proved by Lindenstrauss [25, 5], with the final touch by Soundararajan in the case of the modular surface [36]. They deal with the case of certain hyperbolic surfaces, called arithmetic congruence surfaces; and the eigenfunctions (ϕ_n) are assumed to be common eigenfunctions of Δ and of the Hecke operators ([36] shows that there is no escape of mass to infinity, in the case of noncompact finite volume arithmetic surfaces, such as the modular surface). The methods therein are a very powerful mixture of number theory and ergodic theory. They give, unfortunately, no clue as to the general conjecture.

3.2. Entropy of semiclassical measures on hyperbolic manifolds. The papers [1, 2, 3] deal with the question of quantum unique ergodicity by studying the *Kolmogorov-Sinai entropy* of semiclassical measures. This entropy, denoted h_{KS} in this paper, is a functional going from the set $\mathcal{M}^1_\Phi(S^*X)$ of Φ^t-invariant probability measures on S^*X, to $\mathbb{R}_+$. The shortest definition of the entropy results from a theorem due to Brin and Katok [7]. For any time $T > 0$, introduce a distance on S^*X,

$$d_T(\rho, \rho') = \max_{t\in[-T/2,T/2]} d(\Phi^t\rho, \Phi^t\rho'),$$

where d is the distance built from the Riemannian metric. For $\epsilon > 0$, denote by $B_T(\rho, \epsilon)$ the ball of centre ρ and radius ϵ for the distance d_T. When ϵ is fixed and T goes to infinity, it looks like a thinner and thinner tubular neighbourhood of the geodesic segment $[g^{-\epsilon}\rho, g^{+\epsilon}\rho]$ (this tubular neighbourhood is of radius $e^{-T/2}$ if the curvature of X is constant and equal to -1).

Let μ be a Φ^t–invariant probability measure on S^*X. Then, for μ-almost every ρ, the limit

$$\lim_{\epsilon\longrightarrow 0} \liminf_{T\longrightarrow+\infty} -\frac{1}{T}\log\mu\big(B_T(\rho,\epsilon)\big) = \lim_{\epsilon\longrightarrow 0} \limsup_{T\longrightarrow+\infty} -\frac{1}{T}\log\mu\big(B_T(\rho,\epsilon)\big) \stackrel{\text{def}}{=} h_{KS}(\mu,\rho)$$

exists and it is called the local entropy of the measure μ at the point ρ (it is independent of ρ if μ is ergodic). The Kolmogorov-Sinai entropy is the average of the local entropies: $h_{KS}(\mu) = \int h_{KS}(\mu,\rho)d\mu(\rho)$.

We recall the following (non obvious) facts:

- if $\mu \in \mathcal{M}^1_\Phi(S^*X)$ is carried by a periodic trajectory of Φ^t, then $h_{KS}(\mu) = 0$.
- for all $\mu \in \mathcal{M}^1_\Phi(S^*X)$, we have $0 \le h_{KS}(\mu) \le \int_{S^*X} \sum_{j=1}^{d-1} \lambda_j^+(\rho)d\mu(\rho)$, where the numbers $\lambda_j^+(\rho)$ are the nonnegative Lyapunov exponents of $\rho \in S^*X$ for the geodesic flow (the Ruelle-Pesin inequality). Note that S^*X has dimension $2d-1$. Because the flow is symplectic, there can be at most $d-1$ positive Lyapunov exponents and $d-1$ negative ones.

- If X has negative sectional curvatures, there is equality in the Ruelle-Pesin inequality if and only if μ is the Liouville measure [24].
- the functional h_{KS} is affine.

If the sectional curvature of X is constant equal to -1, the Ruelle-Pesin inequality takes the simpler form: $h_{KS}(\mu) \leq d-1$, with equality if and only if μ is the Liouville measure.

The assumption on the curvature implies that the action of (Φ^t) on S^*X is (uniformly) hyperbolic. This means that, for any $\rho \in S^*X$, the tangent space to S^*X at ρ splits into flow direction, unstable and stable subspaces: there exist $C, \lambda > 0$, and at each $\rho \in S^*X$ a splitting $T_\rho(S^*X) = \mathbb{R}Y(\rho) \oplus E^+_\rho \oplus E^-_\rho$, $\dim E^\pm_\rho = d-1$, such that

(i) For all $\rho \in S^*X$, $d\Phi^t_\rho E^\pm_\rho = E^\pm_{\Phi^t(\rho)}$ for all $t \in \mathbb{R}$;

(ii) For all $\rho \in S^*X$, for all $v \in E^\mp_\rho$, $\|d\Phi^t_\rho.v\| \leq Ce^{-\lambda|t|}\|v\|$, for $\pm t > 0$.

Uniform hyperbolicity is a very strong, and very well understood, form of "chaos".

Let us define the unstable jacobian by

$$\exp \Lambda^+_t(\rho) = \det(d\Phi^t\rceil_{E^+_\rho});$$

for t large enough, we have $\Lambda^+_t(\rho) > 0$ for all ρ.

The following form of Theorem 2.2 is used in [1, 2]. Fix $\delta > 0$, and consider a finite family $\chi_1, \ldots, \chi_K$ of smooth compactly supported functions on T^*X, such that $\sum_{j=1}^K \chi_j \equiv 1$ on $H^{-1}[\frac{1}{2}-\delta, \frac{1}{2}+\delta]$. For all j, assume the function χ_j is supported on a set W_j of diameter $\leq \varepsilon$ (that will be chosen small enough). Also assume that each χ_j vanishes outside $H^{-1}[\frac{1}{2}-2\delta, \frac{1}{2}+2\delta]$. Consider the associated pseudodifferential operators, defined by the Weyl calculus in local coordinates: $\hat{\chi}_j = \chi_j(x, -i\hbar\partial_x)$. Define $\hat{P}_j = e^{\frac{i\tau\hbar\Delta}{2}}\hat{\chi}_j$, for some fixed time step $\tau > 0$. The following theorem amounts to Theorem 2.2 if one works in adapted coordinates in each set W_j:

Theorem 3.3. *In the definition of $\hat{\chi}_j$, we can fix ε, δ small enough, so that the following holds.*

Fix $\mathcal{K} > 0$ arbitrary. Then there exists $\hbar_{\mathcal{K}} > 0$ such that, for $n = \mathcal{K}|\log\hbar|$, for any sequence $(\alpha_1, \ldots, \alpha_n) \in \{1, \ldots, K\}^n$, and for all $\hbar < \hbar_{\mathcal{K}}$,

$$\|\hat{P}_{\alpha_n} \circ \ldots \circ \hat{P}_{\alpha_2} \circ \hat{P}_{\alpha_1}\|_{L^2 \longrightarrow L^2} \leq \frac{1}{(2\pi\hbar)^{d/2}} \prod_{j=1}^n e^{\frac{S_\tau(W_{\alpha_j})}{2}},$$

where $S_\tau(W_j) = -\inf_{\rho \in W_j} \Lambda^+_\tau(\rho)$.

If τ is chosen large enough, then the hyperbolicity condition implies that $S_\tau(W_j) < 0$, and that $\prod_{j=1}^n e^{\frac{S_\tau(W_{\alpha_j})}{2}}$ decays exponentially with n. If the sectional curvature of X is constant, equal to -1, the estimate takes a simpler form:

Theorem 3.4. *Assume that the sectional curvature of X is constant, equal to -1. In the definition of $\hat{\chi}_j$, we can fix ε, δ small enough, so that the following holds.*

Fix $\mathcal{K} > 0$ arbitrary. Then there exists $\hbar_{\mathcal{K}} > 0$ such that, for $n = \mathcal{K}|\log \hbar|$, for any sequence $(\alpha_1, \ldots, \alpha_n) \in \{1, \ldots, K\}^n$, and for all $\hbar < \hbar_{\mathcal{K}}$,

$$\|\hat{P}_{\alpha_n} \circ \ldots \circ \hat{P}_{\alpha_2} \circ \hat{P}_{\alpha_1}\|_{L^2 \longrightarrow L^2} \leq \frac{1}{(2\pi\hbar)^{d/2}} e^{-\left(\frac{d-1}{2}\right)n} (1 + \mathcal{O}(\delta))^n.$$

In [1, 2, 3], we showed how these estimates imply the following lower bound on the entropy of semiclassical measures.

Theorem 3.5. *Let X be a compact d-dimensional riemannian manifold, with negative sectional curvatures. Let (ϕ_λ) be a family of normalized eigenfunctions of the laplacian, $\Delta\phi_\lambda = -\lambda\phi_\lambda$, with $\lambda \longrightarrow +\infty$, and let μ be an associated semiclassical measure. Then:*

[1] We have $h_{KS}(\mu) > 0$.

[2] If the sectional curvature of X is constant, equal to -1, we have $h_{KS}(\mu) \geq \frac{d-1}{2}$.

Remark. *In the case of arithmetic congruence surfaces; and assuming the eigenfunctions (ϕ_λ) are common eigenfunctions of Δ and of the Hecke operators, Bourgain and Lindenstrauss [5] proved the following bound on the measures μ_λ: for any ρ, and all $\epsilon > 0$ small enough,*

$$\mu_\lambda\big(B_T(\rho, \epsilon)\big) \leq Ce^{-T/9}, \tag{6}$$

where the constant C does not depend on ρ or λ. This immediately yields that any semiclassical measure associated with these eigenmodes satisfies $\mu(B_T(\rho, \epsilon)) \leq Ce^{-T/9}$, which implies that any ergodic component of μ has entropy $\geq \frac{1}{9}$. The measure classification result of [25] then implies that μ has to be the Liouville measure.

In [2], Theorem 3.4 is used to prove an estimate that can, in a non rigourous but intuitive manner, be formulated as follows:

$$\mu_\lambda\big(B_T(\rho, \epsilon)\big) \leq C\, \lambda^{\frac{d-1}{4}}\, e^{-\frac{(d-1)T}{2}}. \tag{7}$$

This bound only becomes non-trivial for times $T \gg \log \lambda$. For this reason, we cannot directly deduce bounds on the weights $\mu(B_T(\rho, \epsilon))$; the link between (7) and the entropic bounds of Theorem 3.5 is less direct and uses some specific features of quantum mechanics.

By the properties of entropy, our Theorem 3.5 implies:

Corollary 3.6. *Under the same assumptions,*

[1] If X has (variable) negative sectional curvature, and if γ is a periodic trajectory of Φ^t, then $\mu(\gamma) < 1$.

[2] If the sectional curvature of X is constant, equal to -1, and if γ is a periodic trajectory of Φ^t, then $\mu(\gamma) \leq \frac{1}{2}$.

Corollary 3.7. *[1] If the sectional curvature of X is constant, equal to -1, then the Hausdorff dimension of the support of μ is $\geq d$.*

If X has (variable) negative sectional curvature, we conjectured the following explicit bound for any semiclassical measure μ:

$$h_{KS}(\mu) \geq \frac{1}{2} \int_{S^*X} \sum_{j=1}^{d-1} \lambda_j^+(x,\xi) d\mu(x,\xi). \tag{8}$$

However, in variable curvature, we were not able to push our method that far. This inequality has been proved in the case $d = 2$ by G. Rivière, who was able to extend the proof to nonpositively curved surfaces [29, 30]. In this case, the inequality implies that μ cannot be entirely concentrated on an exponentially unstable closed geodesic.

Proving the quantum unique ergodicity conjecture would be equivalent to getting rid of the $\frac{1}{2}$ factor in (8). This is still far from reach, and would require some new insight into the problem, as there exists an example of a discrete time quantum dynamical system (namely, the "quantum cat-map" [16]) for which equality is reached in (8). This example, however, comes from a symplectic map that is not hamiltonian; see [16] for details.

At the moment, it is not known how to prove (8) when E^+ or E^- have dimension $d - 1 > 1$; or for general non-uniformly hyperbolic systems. The Bunimovich stadium would be a particularly interesting example: the inequality would imply that μ cannot be entirely concentrated on an exponentially unstable periodic trajectory. It would be also be interesting to prove (8) for systems that have some zero Lyapunov exponents. This is one of the motivations for the following paragraph.

3.3. Generalization to higher rank symmetric spaces of nonpositive curvature.

Let G be a connected semisimple Lie group with finite center, let K be a maximal compact subgroup, and G/K the corresponding symmetric space. Let Γ be a cocompact lattice in G, and $X = \Gamma\backslash G/K$.

Example. *Taking $G = \mathrm{SO}_o(1, d; \mathbb{R})$, $K = \mathrm{SO}(d; \mathbb{R})$, one finds that G/K is the d-dimensional real hyperbolic space, which was already treated in the previous paragraph. In this section, one should keep in mind the case $G = \mathrm{SL}(n; \mathbb{R})$, $K = \mathrm{SO}(n; \mathbb{R})$. For $n = 2$, G/K is again the 2-dimensional real hyperbolic space, but from now on we will mostly be interested in $n \geq 3$.*

We will denote by $\mathfrak{g}$ the Lie algebra of G; it is endowed with the Killing bilinear form, which allows to endow G/K with a riemannian metric. We keep using similar calligraphy for Lie subalgebras of $\mathfrak{g}$.

The spectral problem. We look at the algebra $\mathcal{D}$ of G-invariant differential operators on G/K. As a consequence of the structure of semisimple Lie algebras, it is known that $\mathcal{D}$ is commutative, finitely generated. The number of generators r coincides with the *real rank* of G/K, the dimension of a maximal *flat totally geodesic submanifold*; or with the dimension of $\mathfrak{a}$, a maximal *abelian subalgebra* of $\mathfrak{g}$ contained in $\mathfrak{k}^{\perp}$.

Note that $\mathcal{D}$ always contains the laplacian. If $r = 1$, $\mathcal{D}$ is generated by the laplacian, but we will mostly be interested in the case $r \geq 2$.

Example. *For $G = \mathrm{SL}(n, \mathbb{R}), K = \mathrm{SO}(n, \mathbb{R})$, the subalgebra $\mathfrak{a}$ is the set of diagonal matrices with vanishing trace. We will denote by A the connected subgroup of G generated by $\mathfrak{a}$, it consists of diagonal matrices with determinant 1 and nonnegative entries. The rank is the dimension of $\mathfrak{a}$, $r = n-1$. We denote the Weyl group by W, in this example it is the group of permutation matrices. It acts on $\mathfrak{a}$ (and on its dual $\mathfrak{a}^*$).*

We look at the common eigenfunctions of $\mathcal{D}$ on $X = \Gamma\backslash G/K$. The "eigenvalue" is now an r-dimensional vector. In fact, an eigenfunction of $\mathcal{D}$ generates a spherical irreducible representation of G, and these are naturally parametrized by $\nu \in \mathfrak{a}^*/W$. In what follows, the "eigenvalue" will be parametrized by the spectral parameter $\nu \in \mathfrak{a}^*/W$.

The semiclassical limit (as proposed by Silberman-Venkatesh [34]). It consists in the limit

$$\|\nu\| \longrightarrow +\infty, \qquad \frac{\nu}{\|\nu\|} \longrightarrow \nu_\infty. \tag{9}$$

To keep semiclassical notations, one can define $\hbar = \|\nu\|^{-1}$.

We are again interested in the question of quantum ergodicity, which consists in studying a sequence of L^2-normalized eigenfunctions ϕ_ν, of spectral parameters ν, in the asymptotic regime described above. We want to understand the behaviour of the measures $|\phi_\nu(x)|^2 d\mathrm{Vol}(x)$.

The "classical" dynamical system. Consider the algebra $\mathcal{H}$ of G-invariant smooth hamiltonians (i.e. functions) on the cotangent bundle $T^*(G/K)$, that are polynomial in the fibers of the projection $T^*(G/K) \longrightarrow G/K$. Again by the structure of semisimple Lie algebras, $\mathcal{H}$ is commutative under the Poisson bracket, generated by r functions. The algebra $\mathcal{H}$ always contains the quadratic form associated with the Killing metric. Common energy levels of $\mathcal{H}$ are naturally parametrized by $\nu \in \mathfrak{a}^*/W$. We will denote by $\mathcal{E}_\nu$ the energy layer corresponding to the value ν.

We will restrict our attention to non-singular energy levels, in the sense that the generators of $\mathcal{H}$ must have everywhere independent differentials. This is equivalent to ν not being fixed by any element of W: in this case we will say that ν is regular.

The microlocal lift. The measures $|\phi_\nu(x)|^2 d\mathrm{Vol}(x)$ are defined on X. Just as in (4), we study the distributions $\mu_\nu(a) = \langle \phi_\nu, \mathrm{Op}_\hbar(a)\phi_\nu\rangle$ (with $\hbar = \|\nu\|^{-1}$), $\mu_\nu \in \mathcal{D}'(T^*X)$, which project on X to the measure $|\phi_\nu(x)|^2 d\mathrm{Vol}(x)$.

If $a = H \in \mathcal{H}$, then $\mathrm{Op}_\hbar(H)$ is in $\mathcal{D}$, and the isomorphism $H(-i\hbar\bullet) \longleftrightarrow \mathrm{Op}_\hbar(H)$ is the Harish-Chandra isomorphism between $\mathcal{H}$ and $\mathcal{D}$.

The analogue of Theorem 3.1 reads:

Theorem 3.8. *(i) Given any sequence (ν_n) satisfying* (9)*, one can extract from the sequence (μ_{ν_n}) a converging subsequence in $\mathcal{D}'(T^*X)$.*

We will call limits of such subsequences "semiclassical measures" associated with the family (ϕ_{ν_n}), or also "semiclassical measures in the direction ν_∞".

(ii) Let μ be a semiclassical measure in the direction ν_∞. Then μ is a probability measure, carried by the level set $\mathcal{E}_{\nu_\infty}$.

(iii) In addition, for all $H \in \mathcal{H}$, μ is invariant by the hamiltonian flow (Φ_H^t): we have $(\Phi_H^t)_*\mu = \mu$, for all t.

One can extend the quantum unique ergodicity conjecture to this new situation: is it true that the only semiclassical measure in the direction ν_∞ is the Liouville measure on the energy level $\mathcal{E}_{\nu_\infty}$?

Analogously to (8), I would expect the following inequality to hold, for any semiclassical measure μ and all $H \in \mathcal{H}$:

$$h_{KS}(\mu, \Phi_H^t) \geq \frac{1}{2}\sum_j \lambda_j^+(\Phi_H^t).$$

Here $h_{KS}(\mu, \Phi_H^t)$ is the entropy of μ for the flow generated by H, and the $\lambda_j^+(\Phi_H^t)$ are the nonnegative Lyapunov exponents for that flow (since we are on a homogeneous space, each $\lambda_j^+(\Phi_H^t)$ is a constant function). However, the method of [2], so far, can only be pushed to prove the bound:

$$h_{KS}(\mu, \Phi_H^t) \geq \sum_j \left(\lambda_j^+(\Phi_H^t) - \frac{\lambda_{\max}(\Phi_H^t)}{2}\right), \tag{10}$$

where the sum is over all j, and $\lambda_{\max}(\Phi_H^t)$ is the largest of the Lyapunov exponents $\lambda_j^+(\Phi_H^t)$. The right-hand side is, in general, negative, and the lower bound is trivial.

In [4], we are able to prove an explicit, non-trivial lower bound. To do so, we need to get rid of the low Lyapunov exponents in (10). This is where we use the refined norm estimate Theorem 2.3. From now on, we assume that ν_∞ regular.

Theorem 3.9. *[4] Let μ be a semiclassical measure associated to the limit* (9)*. Assume that ν_∞ regular.*

For any $H \in \mathcal{H}$,

$$h_{KS}(\mu, \Phi_H^t) \geq \sum_{j, \lambda_j^+(\Phi_H^t) \geq \frac{\lambda_{\max}(\Phi_H^t)}{2}} \left(\lambda_j^+(\Phi_H^t) - \frac{\lambda_{\max}(\Phi_H^t)}{2}\right).$$

We note that, unless H is a constant function, the entropy lower bound given by Theorem 3.9 is always positive.

One reason to study this problem is that, when the rank r is ≥ 2, the commuting flows (Φ_H^t) $(H \in \mathcal{H})$ are expected to have few joint invariant measures. As a consequence, quantum unique ergodicity should be easier to prove.

To explain what is known about the joint invariant measures of the family (Φ_H^t), we translate everything from the language of hamiltonian flows to the language of group actions. For simplicity we stick to the case $G = \mathrm{SL}(n, \mathbb{R}), K = \mathrm{SO}(n, \mathbb{R})$. If $\mathcal{E}_{\nu_\infty} \subset T^*X$ is a regular energy level of $\mathcal{H}$, it is known that there is a G-equivariant identification between $\mathcal{E}_{\nu_\infty}$ and $\Gamma\backslash G/M$, where M is the group of diagonal matrices of determinant 1 and entries ± 1. Under this identification, the action of the flows $(\Phi_H^t)_{H\in\mathcal{H}}$ on $\mathcal{E}_{\nu_\infty}$ is transported to the right action of the group A on $\Gamma\backslash G/M$. More precisely, if $H \in \mathcal{H}$ is seen as a polynomial function on $\mathfrak{g}^*$, the hamiltonian flow (Φ_H^t) is transported to the 1-parameter subgroup e^{tZ} of A, with $Z = dH(\nu_\infty) \in \mathfrak{a}$ (see [21] for a detailed proof of this fact). In particular, a semiclassical measure μ can be seen as a probability measure on $\Gamma\backslash G/M$, invariant under the right-action of A (in [34], Silberman-Venkatesh constructed a microlocal lift of ϕ_ν, that is directly defined on $\Gamma\backslash G/M$ instead of T^*X, and their construction has the advantage of being equivariant). The Liouville measure on $\mathcal{E}_{\nu_\infty}$ corresponds to the Haar measure on $\Gamma\backslash G/M$.

Margulis' conjecture (see [23]): Let G be a semisimple Lie group with finite center, $\Gamma < G$ a lattice, $A < G$ a maximal split torus. Let μ be an A-invariant and ergodic Borel probability measure on $\Gamma\backslash G$. Then there exists a subgroup L of G, containing A, closed and connected, and a closed orbit $xL \subset \Gamma\backslash G$, such that μ is supported on xL. Also, except possibly when L has a factor of rank 1, μ is **algebraic**, that is the L-invariant measure on xL.

Here is what is known about this conjecture. Let us denote μ_{Haar} the Haar measure on $\Gamma\backslash G$.

- [14], Theorem 4.1 : Let G be an $\mathbb{R}$-split simple group. There exists $0 < c < 1$ such that, if Γ is a lattice of G, and if μ is an A-invariant and ergodic probability measure on $\Gamma\backslash G$ satisfying $h_{KS}(\mu) \geq c\, h_{KS}(\mu_{Haar})$ for every 1-parameter subgroup of A, then μ is the Haar measure on $\Gamma\backslash G$.

  In the case $G = \mathrm{SL}(n, \mathbb{R})$: if μ has positive entropy for each 1-parameter subgroup of A, then μ is the Haar measure on $\Gamma\backslash G$.

- [15] If $G = \mathrm{SL}(n, \mathbb{R})$, $\Gamma = \mathrm{SL}(n, \mathbb{Z})$, or if Γ is a lattice of "inner type"; if μ is ergodic and has positive entropy for some 1-parameter subgroup of A, then μ is algebraic.

- [26, 37] In the latter case, L must be of a certain form : it must be conjugate, via a permutation matrix, to the connected component of identity

in $\mathrm{GL}(t,\mathbb{R})^s \cap \mathrm{SL}(n,\mathbb{R})$; where $n = ts$ and $\mathrm{GL}(t,\mathbb{R})^s$ denotes the block-diagonal embedding of s copies of $\mathrm{GL}(t,\mathbb{R})$ into $\mathrm{GL}(n,\mathbb{R})$.

Here is a reformulation of Theorem 3.9 :

Theorem 3.10. *Let μ be a semiclassical measure associated to the limit* (9)*. Assume that ν_∞ regular.*

Then for any 1-parameter flow e^{tZ} in A (with $Z \in \mathfrak{a}$),

$$h_{KS}(\mu, e^{tZ}) \geq \sum_{j, \lambda_j^+(e^{tZ}) \geq \frac{\lambda_{\max}(e^{tZ})}{2}} \left(\lambda_j^+(e^{tZ}) - \frac{\lambda_{\max}(e^{tZ})}{2} \right).$$

Unless $Z = 0$, this lower bound is positive. In the case $G = \mathrm{SL}(n,\mathbb{R})$, if we knew that μ was ergodic, we could deduce from the result of [14] that μ is the Haar measure, and quantum unique ergodicity would be proved. Unfortunately, nothing tells us that μ is ergodic. However, our entropic lower bound is explicit, and we can use the more precise measure classification results listed above, to prove the following :

Theorem 3.11. *[4] Let $G = \mathrm{SL}(3,\mathbb{R})$, and Γ be any cocompact lattice in G. Let μ be a semiclassical measure associated to the limit* (9)*. Assume that ν_∞ is regular.*

Then μ has a Haar component, of weight $\geq \frac{1}{4}$. In other words, there exists an A-invariant probability measure ν on $\Gamma\backslash G/M$, such that

$$\mu = \frac{1}{4}\mu_{Haar} + \frac{3}{4}\nu,$$

where μ_{Haar} denotes the Haar measure on $\Gamma\backslash G/M$.

Theorem 3.12. *[4] Let $G = \mathrm{SL}(n,\mathbb{R})$, with $n \geq 3$, and let Γ be a lattice associated to a division algebra over $\mathbb{Q}$. Let μ be a semiclassical measure associated to the limit* (9)*. Assume that ν_∞ regular.*

Then μ has a Haar component, of weight $\geq \frac{n-1}{n-d}\left(\frac{1}{2} - \frac{d-1}{n-1}\right)$, where d is the largest proper divisor of n.

We cannot prove quantum unique ergodicity, that says that the only semiclassical measure is the Haar measure. But we have a partial result, saying that any semiclassical measure has a Haar component. For $n = 3$ the result holds for any lattice Γ in $\mathrm{SL}(n,\mathbb{R})$, whereas for $n \geq 4$ we need to assume that Γ is associated to a division algebra over $\mathbb{Q}$ to apply the results of [15, 26, 37].

As a comparison, for n prime, Γ coming from a division algebra over $\mathbb{Q}$, and assuming that the ϕ_ν were also eigenfunctions of the Hecke operators, Silberman-Venkatesh [34, 35] generalized the inequality (6), and improved it by estimating the measures of tubular neighbourhoods of orbits of subgroups. If μ is a semiclassical measure associated to a regular direction ν_∞, their result implies that every ergodic component of μ has positive entropy, with respect to all 1-parameter subgroups of A. This generalizes the result of [5], and implies that μ is the Haar measure.

4. Resonances, Local Smoothing and Strichartz Estimates

Nonnenmacher and Zworski used a variant of Theorem 2.2 in order to prove spectral estimates in scattering theory [27]. For simplicity, we just state their results in a special case. On $\mathbb{R}^d$, consider a Schrödinger operator of the form

$$P(\hbar) = -\hbar^2 \frac{\Delta}{2} + V(x), \quad V \in \mathcal{C}_c^\infty(\mathbb{R}^d, \mathbb{R}),$$

where Δ is the euclidean laplacian. The resonances of $P(\hbar)$ are defined as poles of the meromorphic continuation of the resolvent

$$R(z,\hbar) \stackrel{\text{def}}{=} (P(\hbar) - z)^{-1} : L^2(\mathbb{R}^d) \longrightarrow L^2(\mathbb{R}^d), \quad \Im m(z) > 0,$$

through the continuous spectrum $[0, +\infty)$. More precisely,

$$R(z,\hbar) : L^2_{comp}(\mathbb{R}^d) \longrightarrow L^2_{loc}(\mathbb{R}^d), \quad z \in \mathbb{C} \setminus (-\infty, 0],$$

is a meromorphic family of operators (here L^2_{comp} and L^2_{loc} denote functions which are compactly supported and in L^2, and functions which are locally in L^2). The poles are called resonances, and their set is denoted by $\mathrm{Res}(P(\hbar))$. They are counted according to their multiplicities.

The classical hamiltonian flow is given by Newton's equations :

$$\Phi^t(x,\xi) \stackrel{\text{def}}{=} (x(t), \xi(t)),$$

$$\dot{x}(t) = \xi(t), \dot{\xi}(t) = -dV(x(t)), x(0) = x, \xi(0) = \xi.$$

We will denote $Y = Y_H = \frac{d\Phi^t}{dt}_{t=0}$ the corresponding vector field. This flow preserves the classical hamiltonian

$$H(x,\xi) \stackrel{\text{def}}{=} \frac{\|\xi\|^2}{2} + V(x), \quad (x,\xi) \in \mathbb{R}^d \times \mathbb{R}^d,$$

and it leaves invariant the level sets $\mathcal{E}_E = H^{-1}(E)$. The incoming and outgoing sets at energy E are defined as

$$\Gamma_E^\pm = \{\rho \in \mathcal{E}_E, \Phi^t(\rho) \not\longrightarrow \infty, t \longrightarrow \mp\infty\}.$$

The trapped set at energy E is $K_E = \Gamma_E^+ \cap \Gamma_E^-$. It is a compact invariant set for Φ^t. We will always assume that K_E is non empty.

The fundamental assumption in [27] is that K_E contains no fixed points of the flow, and that the dynamics of (Φ^t) on K_E is (uniformly) hyperbolic. This means that, for any $\rho \in K_E$, the tangent space to $\mathcal{E}_E$ at ρ splits into flow direction, unstable and stable subspaces: there exist $C, \lambda > 0$, and at each

$\rho \in K_E$ a splitting $T_\rho \mathcal{E}_E = \mathbb{R}Y(\rho) \oplus E^+_\rho \oplus E^-_\rho$, $\quad \dim E^\pm_\rho = d-1$, such that

(i) For all $\rho \in K_E$, $d\Phi^t_\rho E^\pm_\rho = E^\pm_{\Phi^t(\rho)}$ for all $t \in \mathbb{R}$;

(ii) For all $\rho \in K_E$, for all $v \in E^\mp_\rho$, $\|d\Phi^t_\rho.v\| \leq Ce^{-\lambda|t|}\|v\|$, for $\pm t > 0$.

Hyperbolicity implies structural stability, and in particular $K_{E'}$ is also a non empty hyperbolic set, for E' close enough to E.

Let us introduce the unstable jacobian, defined by

$$\exp \Lambda^+_t(\rho) = \det(d\Phi^t \rceil_{E^+_\rho});$$

for t large enough, we have $\Lambda^+_t(\rho) > 0$ for all ρ.

By assumption, there exists $R > 0$ such that V is supported inside the ball $B(0,R)$. Fix $\delta > 0$. The technique of complex scaling, used in [27] (but which we don't explain in detail here), allows to construct a deformation $P_\theta(\hbar)$ of $P(\hbar)$ with the following properties : (i) $P_\theta(\hbar)$ is a non self-adjoint deformation of $P(\hbar)$, such that the propagator $e^{-it\frac{P_\theta(\hbar)}{\hbar}}$ damps very rapidly the functions supported away from $B(0,3R)$; (ii) $P_\theta(\hbar)$ coincides with $P(\hbar)$ inside $B(0,2R)$; (iii) the resonances of $P(\hbar)$ close to the real axis are the eigenvalues of $P_\theta(\hbar)$, with the same multiplicities.

The following form of Theorem 2.2 is used in [27]. With the same $\delta > 0$ as previously, consider a finite family $\chi_1, \ldots, \chi_K$ of smooth compactly supported functions on $\mathbb{R}^d \times \mathbb{R}^d$, such that $\sum_{j=1}^K \chi_j \equiv 1$ on $H^{-1}[E-\delta, E+\delta] \cap T^*B(0,R)$. For all j, assume the function χ_j is supported on a set W_j of diameter $\leq \varepsilon$ (that will be chosen small enough). Also assume that each χ_j vanishes outside $H^{-1}[E-2\delta, E+2\delta] \cap T^*B(0,2R)$. Consider the associated pseudodifferential operators, defined by the Weyl calculus : $\hat{\chi}_j = \chi_j(x, -i\hbar\partial_x)$. Define $\hat{P}_j = e^{-i\tau\frac{P(\hbar)}{\hbar}}\hat{\chi}_j$, for some fixed time step $\tau > 0$ (in this definition, it is indifferent to take $P(\hbar)$ or $P_\theta(\hbar)$, since they coincide inside $B(0,2R)$). The following theorem is a variant of Theorem 2.2.

Theorem 4.1. *In the definition of $\hat{\chi}_j$, we can fix ε, δ small enough, so that the following holds.*

Fix $\mathcal{K} > 0$ arbitrary. Then there exists $\hbar_\mathcal{K} > 0$ such that, for $n = \mathcal{K}|\log\hbar|$, for any sequence $(\alpha_1, \ldots, \alpha_n) \in \{1, \ldots, K\}^n$, and for all $\hbar < \hbar_\mathcal{K}$,

$$\|\hat{P}_{\alpha_n} \circ \ldots \circ \hat{P}_{\alpha_2} \circ \hat{P}_{\alpha_1}\|_{L^2 \longrightarrow L^2} \leq \frac{1}{(2\pi\hbar)^{d/2}} \prod_{j=1}^n e^{\frac{S_\tau(W_{\alpha_j})}{2}},$$

where $S_\tau(W_j) = -\inf_{|E'-E|\leq 2\delta, \rho \in W_j \cap K_{E'}} \Lambda^+_\tau(\rho)$.

If τ is chosen large enough, then the hyperbolicity condition implies that $S_\tau(W_j) < 0$, and that $\prod_{j=1}^n e^{\frac{S_\tau(W_{\alpha_j})}{2}}$ decays exponentially with n.

To study the spectral theory of $P_\theta(\hbar)$, we can write

$$e^{-in\tau\frac{P_\theta(\hbar)}{\hbar}}$$
$$= \sum_{(\alpha_1,\dots,\alpha_n)} \hat{P}_{\alpha_n}\circ\dots\circ\hat{P}_{\alpha_2}\circ\hat{P}_{\alpha_1} + \left(e^{-in\tau\frac{P_\theta(\hbar)}{\hbar}} - \sum_{(\alpha_1,\dots,\alpha_n)} \hat{P}_{\alpha_n}\circ\dots\circ\hat{P}_{\alpha_2}\circ\hat{P}_{\alpha_1}\right),$$

where the sum runs over all $(\alpha_1,\dots,\alpha_n)\in\{1,\dots,K\}^n$.

The term $\left(e^{-in\tau\frac{P_\theta(\hbar)}{\hbar}} - \sum_{(\alpha_1,\dots,\alpha_n)} \hat{P}_{\alpha_n}\circ\dots\circ\hat{P}_{\alpha_2}\circ\hat{P}_{\alpha_1}\right)$ only takes into account classical trajectories that, at some time, exit $H^{-1}[E-\delta,E+\delta]\cap T^*B(0,R)$. The trajectories that start inside $H^{-1}[E-\delta,E+\delta]\cap T^*B(0,R)$, but later exit that set, are very rapidly damped by $e^{-in\tau\frac{P_\theta(\hbar)}{\hbar}}$; an important part of [27] is devoted to showing that this term is not relevant when one wants to study the resonance spectrum near $\{\Re e(z)=E\}$. Concerning the other term, we know that each operator $\hat{P}_{\alpha_n}\circ\dots\circ\hat{P}_{\alpha_2}\circ\hat{P}_{\alpha_1}$ has a norm that decays exponentially fast with n, but on the other hand there is an exponential number of terms in the sum $\sum_{(\alpha_1,\dots,\alpha_n)}$. To measure the competition between the exponential number of terms, and the exponential decay of each term, it is natural to introduce the following quantity

$$\mathcal{P}_E(s) = \lim_{\delta\longrightarrow 0}\lim_{\varepsilon\longrightarrow 0}\lim_{n\longrightarrow+\infty}\frac{1}{n\tau}\log Z_{n\tau}(s,(W_j)),$$

where

$$Z_{n\tau}(s,(W_j)) = \inf_B\left\{\sum_{(\alpha_1,\dots,\alpha_n)\in B}\prod_{j=1}^n e^{\frac{S_\tau(W_{\alpha_j})}{2}}\right\},$$

and the inf is taken over all $B\subset\{1,\dots,K\}^n$, such that $K_{E'}\subset\cup_{(\alpha_1,\dots,\alpha_n)\in B}W_{\alpha_1}\cap\Phi^{-\tau}W_{\alpha_2}\cap\dots\cap\Phi^{-(n-1)\tau}W_{\alpha_n}$ for $|E'-E|\leq\delta$.

The function $s\mapsto\mathcal{P}_E(s)$ is called the topological pressure associated with the unstable jacobian. It is strictly decreasing with s.

Corollary 4.2. *Fix $\eta>0$ arbitrary. Then we can find $\tau>0$ large enough, ε,δ small enough, and a partition of unity (χ_j) satisfying all the conditions above, such that the following holds.*

For $\mathcal{K}>0$ arbitrary, there exists $\hbar_{\mathcal{K}}>0$ such that, for $n=\mathcal{K}|\log\hbar|$, and for all $\hbar<\hbar_{\mathcal{K}}$,

$$\left\|\sum_{(\alpha_1,\dots,\alpha_n)}\hat{P}_{\alpha_n}\circ\dots\circ\hat{P}_{\alpha_2}\circ\hat{P}_{\alpha_1}\right\|_{L^2\longrightarrow L^2}\leq\frac{1}{(2\pi\hbar)^{d/2}}e^{n\tau\mathcal{P}_E(\frac{1}{2})}(1+\eta)^{n\tau}.$$

We see that this upper bound is non trivial only if $\mathcal{P}_E(\frac{1}{2})<0$, which means in some sense that the trapped set K_E is rather small. In dimension $d=2$, this

condition is equivalent to saying that the Hausdorff dimension of the trapped set is < 2.

One of the main results in [27] is to deduce from Corollary 4.2 the existence of a spectral gap in the resonance spectrum :

Theorem 4.3. *[27] Assume that $\mathcal{P}_E(\frac{1}{2}) < 0$.*
Then there exists $\delta > 0$ such that, for any γ satisfying

$$0 < \gamma < \min_{|E'-E|\leq\delta} \left(-\mathcal{P}_{E'} \left(\frac{1}{2} \right) \right),$$

there exists $\hbar_{\delta,\gamma} > 0$ such that

$$0 < \hbar < \hbar_{\delta,\gamma} \Longrightarrow \mathrm{Res}(P(\hbar)) \cap ([E-\delta, E+\delta] - i[0, \hbar\gamma]) = \emptyset.$$

This means that if the trapped set is small enough, the resonances stay away from the real axis. This question has been present in the physics literature at least since the seminal paper by Gaspard and Rice [17]. We note that the analogous result for scattering by a disjoint union of convex obstacles was proved in 1988 by Ikawa [22]. One can say that Ikawa's paper contained, in a hidden form and in a specific geometric situation, the idea expressed by Theorem 2.1.

One important consequence of Corollary 4.2 is the following estimate on the resolvent. It is proved in [27], using the relation between the resolvent and the propagator.

Theorem 4.4. *[27] Assume that $\mathcal{P}_E(\frac{1}{2}) < 0$. Then, for any $\chi \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, there exists $C > 0$ such that*

$$\|\chi(P(\hbar) - E)^{-1}\chi\|_{L^2 \longrightarrow L^2} \leq \frac{C \log|\hbar|}{\hbar},$$

for $\hbar$ small enough.

These theorems hold for more general operators : see [27] for a more general set of assumptions. An interesting situation is when there is no potential ($V = 0$) and one studies the resonance spectrum of the laplacian for a riemannian metric that is euclidean ouside a compact set[1]. Since the hamiltonian is homogeneous, one can without loss of generality consider the case $E = \frac{1}{2}$, that is, our hamiltonian flow is the unit geodesic flow. In this situation, the resolvent estimate above was extended by Datchev to the case of asymptotically conic manifolds, also called scattering manifolds [12]. It is shown in [9, 12] how such resolvent estimates imply a local smoothing estimate :

Theorem 4.5. *[27, 9, 12] Let (X, g) be a riemannian manifold that is euclidean outside a compact set; or asymptotically conic (see [12] for the definition). Let Δ denote the associated Laplace-Beltrami operator.*

[1] In the case of convex-cocompact hyperbolic manifolds, the existence of a gap in the resonance spectrum, if the limit set has small dimension, seems to have been known before.

Assume that the trapped set K of the unit speed geodesic flow is compact, hyperbolic, and that the pressure of the unstable jacobian on K satisfies $\mathcal{P}(\frac{1}{2}) < 0$.

Then, for any $\eta > 0$, for any $T > 0$ and any $\chi \in \mathcal{C}_c^\infty(M)$, there exists $C > 0$ such that

$$\int_0^T \|\chi e^{it\Delta} u\|^2_{H^{1/2-\eta}} dt \leq C\|u\|^2_{L^2}. \tag{11}$$

The local smoothing effect usually refers to the inequality

$$\int_0^T \|\chi e^{it\Delta} u\|^2_{H^{1/2}} dt \leq C\|u\|^2_{L^2},$$

which is known to hold when the trapped set for the geodesic flow is empty. Doi [13] showed, in a variety of geometric situations, that the absence of trapped geodesics is also a necessary condition for (11) to hold with $\eta = 0$. According to Theorem 4.5, if the trapped set is hyperbolic and small enough, (11) holds for all $\eta > 0$, which is called "local smoothing with loss".

Burq-Guillarmou-Hassell [6] showed how the combination of Theorem 4.5 and the norm estimate of Corollary 4.2 yields a Strichartz estimate *without* loss :

$$\|e^{it\Delta} u\|_{L^p((0,1),L^q(M))} \leq C\|u\|_{L^2(M)},$$

for $\frac{2}{p} + \frac{d}{q} = \frac{d}{2}$, $p > 2, q \geq 2$, $(p,q) \neq (2,\infty)$. This estimate holds for riemannian manifolds that are asymptotically conic, assuming that the trapped set of the unit geodesic flow is compact, hyperbolic and satisfies $\mathcal{P}(\frac{1}{2}) < 0$.

Finally, Christianson [10] and Nonnenmacher-Zworski [28] show how to extend the resolvent estimate of Theorem 4.4 to the analytic extension of the cut-off resolvent in a small strip below the real axis. As an application, Christianson [10] proves exponential decay of the local energy, under the action of the wave group, on a riemannian manifold that it euclidean outside a compact set, assuming once again that the trapped set of the unit geodesic flow is hyperbolic and satisfies $\mathcal{P}(\frac{1}{2}) < 0$.

We refer the reader to the work of Emmanuel Schenck [32], who used similar ideas to study the spectrum and the energy decay for the damped wave equation.

References

[1] N. Anantharaman, *Entropy and the localization of eigenfunctions*, Ann. of Math. (2) 168 (2008), no. 2, 435–475.

[2] N. Anantharaman, S. Nonnenmacher, *Half-delocalization of eigenfunctions for the laplacian on an Anosov manifold*, Festival Yves Colin de Verdière. Ann. Inst. Fourier (Grenoble) 57 (2007), no. 7, 2465–2523.

[3] N. Anantharaman, H. Koch and S. Nonnenmacher, *Entropy of eigenfunctions*, to appear in the Proceedings of the International Congress on Mathematical Physics - Rio de Janeiro, August 6–11, 2006.

[4] N. Anantharaman, L. Silberman, *Asymptotic distribution of eigenfunctions on locally symmetric spaces*, work in progress.

[5] J. Bourgain, E. Lindenstrauss, *Entropy of quantum limits,* Comm. Math. Phys. 233 (2003), 153–171.

[6] N. Burq, C. Guillarmou, A. Hassell, *Strichartz estimates without loss on manifolds with hyperbolic trapped geodesics*, preprint.

[7] M. Brin, A. Katok, *On local entropy,* Geometric dynamics (Rio de Janeiro, 1981), 30–38, Lecture Notes in Math., 1007, Springer, Berlin, 1983.

[8] H. Christianson, *Semiclassical non-concentration near hyperbolic orbits*, J. Funct. Anal. 262 (2007), 145–195; ibid, *Dispersive estimates for manifolds with one trapped orbit,* Comm. PDE 33 (2008), 1147–1174.

[9] H. Christianson, *Cutoff resolvent estimates and the semilinear Schrödinger equation*, Proc. AMS 136 (2008), 3513–3520.

[10] H. Christianson, *Applications to cut-off resolvent estimates to the wave equation*, Math. Res. Lett. Vol. 16 (2009), no. 4, 577–590.

[11] Y. Colin de Verdière, *Ergodicité et fonctions propres du laplacien*, Commun. Math. Phys. 102, (1985) 497–502.

[12] K. Datchev, *Local smoothing for scattering manifolds with hyperbolic trapped sets*, Comm. Math. Phys. 286, no. 3, (2009) 837–850.

[13] S.–I. Doi, *Smoothing effects of Schrödinger evolution groups on Riemannian manifolds*, Duke Math. J. 82 (1996), 679–706.

[14] M. Einsiedler, A. Katok, *Invariant measures on G/Γ for split simple Lie groups G.* Dedicated to the memory of Jrgen K. Moser. Comm. Pure Appl. Math. 56 (2003), no. 8, 1184–1221.

[15] M. Einsiedler, A. Katok, E. Lindenstrauss, *Invariant measures and the set of exceptions to Littlewood's conjecture.* Ann. of Math. (2) 164 (2006), no. 2, 513–560.

[16] F. Faure, S. Nonnenmacher and S. De Bièvre, *Scarred eigenstates for quantum cat maps of minimal periods*, Commun. Math. Phys. 239 (2003), 449–492.

[17] P. Gaspard, S.A. Rice, *Semiclassical quantization of the scattering from a classically chaotic repellor,* J. Chem. Phys. 90 (1989), 2242–2254.

[18] P. Gérard, E. Leichtnam, *Ergodic properties of eigenfunctions for the Dirichlet problem.* Duke Math. Jour. 71 (1993), 559–607.

[19] A. Hassell, *Ergodic billiards that are not quantum unique ergodic.* With an appendix by A. Hassell and L. Hillairet. Annals of Mathematics, to appear.

[20] B. Helffer, A. Martinez, D. Robert, *Ergodicité et limite semi-classique.* Comm. in Math. Phys. 109 (1987), 313–326.

[21] J. Hilgert *An ergodic Arnold-Liouville theorem for locally symmetric spaces.* Twenty years of Bialowieza: a mathematical anthology, 163–184, World Sci. Monogr. Ser. Math., 8, World Sci. Publ., Hackensack, NJ, 2005.

[22] M. Ikawa, *Decay of solutions of the wave equation in the exterior of several convex bodies.* Ann. Inst. Fourier, 38 (1988), 113–146.

[23] A. Katok, R. Spatzier, *Invariant measures for higher rank hyperbolic Abelian actions.* Erg. Theory and Dynam. Systems, 16 (1996), 751–778.

[24] F. Ledrappier, L.-S. Young, *The metric entropy of diffeomorphisms. I. Characterization of measures satisfying Pesin's entropy formula*, Ann. of Math. (2) 122 (1985), no. 3, 509–539.

[25] E. Lindenstrauss, *Invariant measures and arithmetic quantum unique ergodicity*, Ann. Math. 163 (2006), 165–219.

[26] E. Lindenstrauss, B. Weiss, *On sets invariant under the action of the diagonal group.* Ergodic Theory Dynam. Systems 21 (2001), no. 5, 1481–1500.

[27] S. Nonnenmacher, M. Zworski, *Quantum decay rates in chaotic scattering*, Acta Mathematica, Volume 203, Number 2, December 2009, 149–233.

[28] S. Nonnenmacher, M. Zworski, *Semiclassical resolvent estimates in chaotic scattering*, Appl. Math. Res. Express (2009), 74–86.

[29] G. Rivière, *Entropy of semiclassical measures in dimension 2*, to appear in Duke Math. J.

[30] G. Rivière, *Entropy of semiclassical measures for nonpositively curved surfaces*, preprint.

[31] Z. Rudnick and P. Sarnak, *The behaviour of eigenstates of arithmetic hyperbolic manifolds*, Commun. Math. Phys. 161 (1994), 195–213.

[32] E. Schenck, *Energy decay for the damped wave equation under a pressure condition*, preprint 2009.

[33] A. Schnirelman, *Ergodic properties of eigenfunctions*, Usp. Math. Nauk. 29 (1974), 181–182.

[34] L. Silberman, A. Venkatesh, *Quantum unique ergodicity for locally symmetric spaces,* to appear in GAFA.

[35] L. Silberman, A. Venkatesh, *Entropy bounds for Hecke eigenfunctions on division algebras,* preprint.

[36] K. Soundararajan, *Quantum unique ergodicity for* $SL_2(\mathbb{Z})\backslash\mathbb{H}$, preprint.

[37] G. Tomanov, *Actions of maximal tori on homogeneous spaces.* Rigidity in dynamics and geometry (Cambridge, 2000), 407–424, Springer, Berlin, 2002.

[38] S. Zelditch, *Uniform distribution of the eigenfunctions on compact hyperbolic surfaces*, Duke Math. J. 55 (1987), 919–941.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Random Data Cauchy Theory for Dispersive Partial Differential Equations

Nicolas Burq*

Abstract

In a series of papers in 1930-32, Paley and Zygmund proved that random series on the torus enjoy better L^p bounds than the bounds predicted by the deterministic approach (and Sobolev embeddings). The subject of random series was later largely studied and developed in the context of harmonic analysis. Curiously, this phenomenon was until recently not exploited in the context of partial differential equations. The purpose of this talk is precisely to present some recent results showing that in some sense, the solutions of dispersive equations such as Schrödinger or wave equations are better behaved when one consider initial data randomly chosen (in some sense) than what would be predicted by the deterministic theory. A large part of the material presented here is a collaboration with N. Tzvetkov.

Mathematics Subject Classification (2010). Primary 35LXX; Secondary 35Q55.

Keywords. Random series, Wave equations, Schrödinger equations

1. Introduction

In a series of papers in 1930-32, Paley and Zygmund [42] proved that for any square summable sequence $(c_n) \in \ell^2$, if one consider the trigonometric series

$$u(x) = \sum_{n=0}^{+\infty} c_n e^{inx},$$

then, changing the signs of the coefficients c_n randomly and independently ensures that almost surely, the sum of the series is in every space $L^p(\mathbb{T})$, $2 \leq p < +\infty$. In modern language, this result reads

*Mathématiques, Bât. 425, Université Paris-Sud 11, 91405 Orsay Cedex, France.
E-mail: nicolas.burq@math.u-psud.fr

Theorem 1.1. *Consider a sequence* $(c_n) \in \ell^2$ *and a family of* independent, mean zero *Bernouilli random variables,* (b_n^ω) *on a probability space* $(\Omega, \mathcal{P})$*:*

$$\mathcal{P}(b_n = \pm 1) = \frac{1}{2},$$

and the corresponding series on the torus,

$$u^\omega(x) = \sum_{n=0}^{+\infty} b_n^\omega c_n e^{inx}.$$

Then almost surely

$$\forall 2 \leq p < +\infty, u^\omega \in L^p(\mathbb{T}).$$

Actually, in 1930, the most difficult part in this result was precisely to define what is a "*family of independent, mean zero Bernouilli random variables*", and Paley-Zygmund proof relied on an explicit realization (see Rademacher [43] and Kolmogorov [27]). With modern technology, it is not difficult to give a quantitative version of this result and one can prove (see section 2)

$$\forall 2 \leq p < +\infty, \exists C > 0; \forall \lambda > 0, \mathcal{P}(\|u^\omega\|_{L^p(\mathbb{T})} > \lambda) \leq Ce^{-\lambda^2/C}.$$

This much celebrated result has been followed by many works on random series of functions (see in particular the books by Kahane [24] and Marcus-Pisier [36]) where the studies focused mostly on the question of giving criteria for the uniform convergence of the series. It is quite remarkable that this very active fields of research for the point of view of harmonic analysis were not, until recently investigated from the point of view of partial differential equations. To my knowledge, the first step toward this direction is due to Bourgain [7], where these properties of random series on the torus $\mathbb{T}^2$ were exploited in the context of the (renormalized by Wick ordering) two dimensional non linear cubic Schrödinger equations. The purpose of this talk is in fact to show that these properties of random series can be exploited in a number of situations including wave equations on manifolds and non linear harmonic oscillators. The examples we have in mind are the semilinear wave equation on a compact manifold

$$\left(\frac{\partial^2}{\partial t^2} - \mathbf{\Delta}\right) u = -|u|^{p-1}u, \qquad u\,|_{t=0} = u_0, \quad \frac{\partial}{\partial t} u\,|_{t=0} = u_1, \tag{1}$$

and the semilinear Schrödinger equation on the line

$$\left(i\frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2} + x^2\right) u = -\kappa|u|^{p-1}u, \qquad u\,|_{t=0} = u_0, \qquad \kappa = 0; \pm 1. \tag{2}$$

As far as Cauchy theory is concerned, the (deterministic) behaviour of these equations has been investigated for a long time and the picture is by now fairly complete. Notice that up to now, the ideas presented in this talk do not apply to the case of nonlinear Schrödinger equations on compact manifolds (see Tzvetkov [47, Appendix] where some partial results are obtained in this case). Notice also that in this setting of deterministic theory of semi-linear Schrödinger equations on manifolds, the situation is much less well known, see Gérard [19] for a review of this question). Many of the questions which remain open on $\mathbb{R}^d$ are essentialy about the critical problems and the long time behaviour (or possibly explosion, see the works by Merle-Raphael [37, 38, 40, 39]) of the solutions. In particular, for both the wave equation on a compact manifold, and the Schrödinger equation, the Cauchy problem is known to be well posed above the scaling index

$$s_c = \frac{d}{2} - \frac{2}{p-1}.$$

(see Kapitanskii [25], Oh [41] and Carles [15], and the contributions by Bourgain [6], Colliander-Keel-Staffilani-Takaoka-Tao [17], and Kenig-Merle [29, 30] for the critical problems), while it is known to be ill posed below the scaling index. Indeed, the following result is known (see the works by Lebeau [32, 33], Christ-Colliander-Tao [16], Burq-Gérard-Tzvetkov [9], Alazard-Carles [1] and Thomann [48]).

Theorem 1.2. *Assume that $0 < s < s_c$. Then there exists a sequence of initial data $(u_{0,n}, u_{1,n}) \to 0$ in $H^s(M) \times H^{s-1}(M)$ as $n \to 0$, and there exists times $t_n \to 0$ such that the solutions of* (1) *exist (and are unique) in suitable spaces for $|t| \leq t_n$, but*

$$\lim_{n\to+\infty} \|u(t_n)\|_{H^s(M)} = +\infty.$$

There exists also a sequence of initial data $(u_{0,n}) \to 0$ in $H^s(\mathbb{R}))$ as $n \to 0$, and there exists times $t_n \to 0$ such that the solutions of (2) *exist (and are unique) in suitable spaces for $|t| \leq t_n$, but*

$$\lim_{n\to+\infty} \|u(t_n)\|_{H^s(\mathbb{R})} = +\infty.$$

In other word, the equations (1) and (2) admit no flow continuous at $t = 0$, $(u_0, u_1) = 0$ (resp $u_0 = 0$) for the H^s topology. Having this negative result in mind, a natural question to ask is whether one can still find initial states with super-critial regularity (i.e. $(u_0, u_1) \in H^s(M) \times H^{s-1}(M)$, (resp. $u_0 \in H^s(M)$), $s < s_c$), for which the Cauchy problem (1) (resp. (2)) is locally (or even better, globally) well posed. The purpose of this talk is precisely to present such examples.

The paper is organized as follows: In Section 2, I will present a short proof of Paley-Zygmund's result which, using Hörmander-Sogge's Laplace eigenfunc-

tions estimates [21, 45], or Hermite eigenfunctions estimates [31] extends readily to the more general setting of random series on manifolds (or on $\mathbb{R}^d$). In section 3, I will show how these estimates, combined with the usual Strichartz estimates [46, 20, 28] allow to obtain a nice "probabilistic" Cauchy theory for the wave equation on compact manifolds and I will give a particular example where this local theory combined with Bourgain's [7, 6] Gibbs measure arguments gives a global (in time) result. In Section 4, I will follow the same program for the semi-linear Schrödinger equation on the line $\mathbb{R}$, with or without harmonic oscillator. Finally, in a last section, I will focus on some different randomizations in connexion with Sobolev embeddings.

2. Random Series

2.1. Random series on the torus.

In this section I will give a simple proof of Paley-Zygmund's theorem, to show the versatility of the result.

Theorem 2.1 (see [2, 13, 14]). *Assume that the random variable b_n^ω are*

1. *independent,*
2. *have mean equal to* 0*,*
3. *have super-exponential decay*

$$\exists C, \delta > 0; \forall \alpha \in \mathbb{R}, \mathbb{E}(e^{\alpha |b_n^\omega|}) \leq C e^{\delta \alpha^2}. \tag{3}$$

Notice that this latter assumption is satisfied for Bernouilli, or more generally for families of random variables having a (fixed) compact support, or for standard Gaussian random variables.

Then, almost surely, $u_n^\omega \in L^p(\mathbb{T}), \forall q < +\infty$. More precisely, the following large deviation estimate holds

$$\forall q < +\infty, \exists C; \qquad \mathcal{P}(\|u^\omega\|_{L^q(\mathbb{T})} > \lambda) \leq C e^{-\Lambda^2/C}.$$

The remaining of this section is devoted to the proof of Theorem 2.1.

2.2. Proof of Theorem 2.1.

The proof relies on

Proposition 2.2. *[Large deviation estimate] Assume that the random variables satisfy the assumptions of Theorem 2.1. Then there exists $\delta > 0$ such that for any $\Lambda > 0$, and any sequence $(v_n) \in \ell^2$,*

$$\mathcal{P}\left(\left| \sum_n v_n b_n^\omega \right| > \Lambda \right) \leq e^{-\delta \frac{\Lambda^2}{\sum_n |v_n|^2}}.$$

2.2.1. Proof of Theorem 2.1 assuming Proposition 2.2. Fix $r \geq q$. Remark that the norm of an integral is always smaller that the integral of the norm. As a consequence,

$$\begin{aligned} \| \|u^\omega(x)\|_{L^q_x} \|_{L^r_\omega} &= \left(\| \int_x |u^\omega(x)|^q dx \|_{L^{r/q}_\omega} \right)^{1/q}, \\ &\leq \left(\int_x \| |u^\omega(x)|^q \|_{L^{r/q}_\omega} dx \right)^{1/q}, \\ &= \| \|u^\omega(x)\|_{L^r_\omega} \|_{L^q_x}. \end{aligned} \tag{4}$$

Notice (x is a fixed parameter) that

$$\|u^\omega(x)\|_{L^r_\omega} = \int_0^{+\infty} r\lambda^{r-1} \mathcal{P}(|u^\omega(x)| > \lambda) d\lambda,$$

and according to Proposition 2.2 applied to $v_n = u_n e^{inx}$, with x a fixed parameter (and the change of variables $\mu = \left(\frac{\sqrt{2}\delta^{1/2}}{\sum_n |u_n e^{inx}|^2)^{1/2}} \right)$,

$$\begin{aligned} \|u^\omega(x)\|^r_{L^r_\omega} &\leq C \int_0^{+\infty} r\lambda^{r-1} e^{-\delta \frac{\lambda^2}{\sum_n |u_n e^{inx}|^2}} d\lambda \\ &\leq \left(C \sum_n |u_n|^2 \right)^{r/2} r \int_0^{+\infty} \mu^{r-1} e^{-\frac{\mu^2}{2}} d\mu, \\ &\leq \left(C \sum_n |u_n|^2 \right)^{r/2} \times r \times r-2 \times \cdots \times 1 \leq \left(C' r \sum_n |u_n|^2 \right)^{r/2}. \end{aligned} \tag{5}$$

Notice now that the norm with respect to the x parameter is harmless (as the bound does not depend on x). For future use, it should be noticed that we use here that the L^q norm of the functions e^{inx} are uniformly bounded. As a conclusion, we just proved

$$\| \|u^\omega(x)\|_{L^q_x} \|_{L^r_\omega} \leq \left(C' r \sum_n |u_n|^2 \right)^{1/2}.$$

To conclude, let us recall Tchebytchev inequality:

$$\forall \lambda, \qquad \lambda \mathcal{P}(f^\omega > \Lambda) \leq \mathbb{E}(f).$$

Apply this inequality to the random variable $f^\omega = \|u^\omega(x)\|^r_{L^q_x}$ and $\lambda = \Lambda^r$. We get

$$\begin{aligned} \mathcal{P}(\|u^\omega(x)\|_{L^q_x} > \Lambda) &= \mathcal{P}(\|u^\omega(x)\|^r_{L^q_x} > \Lambda^r = \lambda) \\ &\leq \frac{1}{\Lambda^r} \mathbb{E}(\|u^\omega(x)\|^r_{L^q_x}) = \frac{1}{\Lambda^r} \| \|u^\omega(x)\|_{L^q_x} \|^r_{L^r_\omega} \\ &\leq \left(\frac{(C' r \sum_n |u_n|^2)}{\Lambda^2} \right)^{r/2}. \end{aligned} \tag{6}$$

Now we optimize this inequality by choosing r so that

$$\frac{(C'r\sum_n |u_n|^2)}{\Lambda^2} = \frac{1}{2}$$

(notice that the assumption $r \geq p$ requires that λ is large enough, but for bounded λ's, the large deviation estimate in Theorem 2.1 is straightforward). This gives

$$\mathcal{P}(\|u^\omega(x)\|_{L^q_x} > \Lambda) \leq \left(\frac{1}{2}\right)^{r/2} = e^{-\delta\frac{\Lambda^2}{\sum_n |u_n|^2}},$$

which ends the proof of Theorem 2.1. □

2.2.2. Proof of Proposition 2.2. The proof we give is very classical. In the special case where the random variables g_n are gaussian random variables of variance 1, the result is straightforward. Indeed, $\sum_n v_n g_n$ is a Gaussian random variable of variance $\sum_n |v_n|^2$ and the result follows. In the general case, it is enough to prove

$$\mathcal{P}\left(\sum_n v_n b_n^\omega > \lambda\right) \leq e^{-\delta\frac{\lambda^2}{\sum_n |v_n|^2}}.$$

Indeed, the estimate for the other part, $\mathcal{P}(\sum_n v_n b_n^\omega < -\lambda)$ is obtained by changing v_n to $-v_n$. Let us fix $t > 0$ and compute (using the fact that the random variables are independent)

$$\begin{aligned}\mathbb{E}\left(e^{t\sum_n v_n b_n^\omega}\right) &= \mathbb{E}\left(\prod_n e^{tv_n b_n^\omega}\right) = \prod_n \mathbb{E}\left(e^{tv_n b_n^\omega}\right)\\ &\leq \prod_n e^{\delta t^2 |v_n|^2} \leq e^{t^2\sum_n |v_n|^2},\end{aligned} \tag{7}$$

where in the last but one inequality, we used the super-exponential decay assumption(3). Now, using Tchebytchev inequality,

$$\begin{aligned}\mathcal{P}\left(\sum_n v_n b_n^\omega > \lambda\right) &= \mathcal{P}\left(e^{t\sum_n v_n b_n^\omega} > e^{t\lambda}\right),\\ &\leq e^{-t\lambda}\mathbb{E}\left(e^{t\sum_n v_n b_n^\omega}\right),\\ &\leq e^{\delta t^2\sum_n |v_n|^2 - t\lambda}.\end{aligned} \tag{8}$$

Optimize by choosing $\delta t^2 \sum_n |v_n|^2 = t\lambda/2$, i.e. $t = \lambda/(2\delta\sum_n |v_n|^2)$, which gives

$$\mathcal{P}\left(\sum_n v_n b_n^\omega > \lambda\right) \leq e^{-\alpha\frac{\lambda^2}{\sum_n |v_n|^2}},$$

which ends the proof of Proposition 2.2 and consequently the proof of Theorem 2.1. □

2.3. Random series on manifolds and on the line. Consider M a riemanian manifold and H a non-negative self adjoint operator on $L^2(M)$ with compact resolvent (the examples we have in mind are M a compact riemanian manifold with $H = -\mathbf{\Delta}$ and $M = \mathbb{R}$ with $H = -\frac{d^2}{dx^2} + x^2$ the harmonic oscillator). It is well known that the eigenfunctions of H, e_n, associated to eigenvalues $-\lambda_n^2$ provide a Hilbert base of $L^2(M)$

$$u \in L^2(M) \Leftrightarrow u = \sum_{n\in\mathbb{N}} u_n e_n(x), \|u\|^2_{L^2(M)} = \sum_{n\in\mathbb{N}} |u_n|^2 < +\infty$$

Definition 1. For any $s \in \mathbb{R}$, let $\mathcal{H}^s(M)$ be the space of distributions such that $(\mathrm{Id}+H)^s u \in L^2(M)$, and let $\mathcal{W}^{s,p}(M)$ be the space of distributions u such that $(\mathrm{Id}+H)^{s/2}u \in L^p(M)$ endowed with their natural norm. In particular, we have

$$u \in \mathcal{H}^s(M) \Leftrightarrow u = \sum_{n\in\mathbb{N}} u_n e_n(x), \sum_{n\in\mathbb{N}} (1+\lambda_n^2)^s)|u_n|^2 = \|u_n\|^2_{\mathcal{H}^s(M)} < +\infty$$

and notice that if M is a compact manifold, $\mathcal{H}^s(M)$ coincides with the usual Sobolev space $H^s(M)$ while if M is the real line and H the harmonic oscillator, and $s \geq 0$

$$\mathcal{W}^{s,p}(M) = \{u \in \mathcal{D}'(\mathbb{R}); \langle x\rangle^s u, \langle |D_x|\rangle^s u \in L^p(\mathbb{R})$$

(endowed with its natural norm [18]).

The starting point of the analysis is Hörmander-Sogge's estimates for the growth of the L^p norm of eigenfunctions on (compact) manifolds

Theorem 2.3. *Consider M a compact riemanian manifold and $(e_n)_{n\in\mathbb{N}}$, the L^2- normalized eigenfunctions of the Laplace operator on M, associated to the eigenvalues $-\lambda_n^2$. Then, there exists $C > 0$ such that for any $n \in \mathbb{N}$, and any $2 \leq p \leq +\infty$*

$$\|e_n\|_{L^p(M)} \leq C\lambda_n^{\delta(p)} \tag{9}$$

where

$$\delta(p) = \begin{cases} \frac{d-1}{2} - \frac{d}{p} \text{ if } p \geq \frac{2(d+1)}{d-1}, \\ \frac{d-1}{2}\left(\frac{1}{2} - \frac{1}{p}\right) \text{ if } p \leq \frac{2(d+1)}{d-1}. \end{cases} \tag{10}$$

The end point $p = \infty$ is due to Hörmander [21] while the point $p = \frac{2(d+1)}{d-1}$ is due to Sogge [45] (notice that the last extremal point $p = 2$ is trivial).

Consider now the (L^2 normalized) eigenfunctions of the Harmonic oscillator, $h_n(x)$,

$$\left(-\frac{d^2}{dx^2}+x^2\right)h_n=\lambda_n^2 h_n, \lambda_n=\sqrt{2n+1}$$

Then the analog of Sogge's result is the following (see Yajima-Zhang [49] and Koch-Tataru [31]

Theorem 2.4. *For any $2\leq p\leq+\infty$, there exists $C>0$ such that for any $n\in\mathbb{N}$,*

$$\|h_n\|_{L^p(\mathbb{R})}\leq C\lambda_n^{\sigma(p)} \tag{11}$$

with

$$\sigma(p)=\begin{cases} -\left(\frac{1}{6}+\frac{1}{3p}\right) \text{ if } 4<p\leq+\infty \\ =-\left(\frac{1}{2}-\frac{1}{p}\right) \text{ if } 2\leq p<4 \end{cases} \tag{12}$$

and

$$\|h_n\|_{L^4(\mathbb{R})}\leq C\lambda_n^{-\frac{1}{4}}\log(\lambda_n)^{1/4} \tag{13}$$

Remark. *Notice that in the case of the harmonic oscillator, the situation is much more favorable as the L^p norms of the Hermite functions h_n tend to be small as n tend to infinity. This is of course natural, as, by elliptic regularity, the functions h_n are essentially concentrated in the set $\{|x|\leq\lambda_n\}$, whose measure is growing.*

Remark. *Following the $X^{s,b}$ approach by Bourgain [4, 3, 5], multilinear versions of estimates* (9) *proved to be crucial in the analysis of the (deterministic) well posedness of non linear Schrödinger equations on general compact manifolds and spheres (see the works by Burq-Gérard-Tzvetkov [8, 10, 19]), while the bilinear version of* (13) *was the starting point of our work on the non linear harmonic oscillator (see [12] and Section 4.1).*

Now the analog of Paley-Zygmund's theorem is (see Burq-Tzvetkov [13])

Theorem 2.5. *Consider a compact riemanian manifold, M. Fix $2\leq p<+\infty$ and consider*

$$u=\sum_n u_n e_n(x)\in\mathcal{H}^s(M),$$

and random variables (b_n) satisfying the assumptions in Theorem 2.1. Assume that $s>\delta(p)$. Then almost surely the random series

$$u^\omega=\sum_{n\in\mathbb{N}} b_n^\omega u_n e_n(x)$$

belongs to $L^p(M)$ and more precisely

$$\exists C>0; \mathcal{P}(\|u^\omega\|_{L^p(M)}>\lambda)\leq Ce^{-\lambda^2/C}. \tag{14}$$

Furthermore, for any $s'>s$, if $u\notin \mathcal{H}^{s'}(M)$, then

$$\mathcal{P}(\|u^\omega\|_{\mathcal{H}^{s'}(M)}<+\infty)=0. \tag{15}$$

In the case of the harmonic oscillator, the analog of Paley-Zygmund's theorem is (see Burq-Thomann-Tzvetkov [12])

Theorem 2.6. *Fix $2\leq p<+\infty$ and consider*

$$u=\sum_n u_n h_n(x)\in \mathcal{H}^s(\mathbb{R}),$$

and random variables (b_n) satisfying the assumptions in Theorem 2.1. Assume that $s>\sigma(p)$. Then almost surely the random series

$$u^\omega=\sum_{n\in\mathbb{N}} b_n^\omega u_n h_n(x)$$

belongs to $L^p(\mathbb{R})$ and more precisely

$$\exists C>0; \mathcal{P}(\|u^\omega\|_{L^p(\mathbb{R})}>\lambda)\leq Ce^{-\lambda^2/C}. \tag{16}$$

Furthermore, for any $s'>s$, if $u\notin \mathcal{H}^{s'}(\mathbb{R})$, then

$$\mathcal{P}(\|u^\omega\|_{\mathcal{H}^{s'}(\mathbb{R})}<+\infty)=0. \tag{17}$$

Remark. *Notice that these results exhibit gains of derivatives with respect to the Sobolev embeddings. Indeed, it is of course clear for the harmonic oscillator case as the L^p norms are better behaved almost surely than the L^2 norms, while in the case of a compact manifold, Sobolev embeddings read*

$$\|u\|_{L^p(M)}\leq C\|u\|_{\mathcal{H}^s(M)},\qquad s=d\left(\frac12-\frac1p\right),\quad 2\leq p<+\infty.$$

3. Wave Equations and Random Series

3.1. Local theory. In this section, for simplicity, I shall consider the simplest model on semi-linear wave equation, which is obtained for cubic non linearities on three dimensional manifolds.

$$(\partial_t^2-\Delta)u+u^3=0,\quad (u,\partial_t u)|_{t=0}=(u_1,u_2)\in H^s(M)\times H^{s-1}(M). \tag{18}$$

Notice that for this equation, the critical index is $s_c = \frac{1}{2}$. The following result (Burq-Tzvetkov [13]) shows that neverthless, the Cauchy problem is locally well posed for a large number of supercritical initial data

Theorem 3.1. *Consider a compact riemanian manifold, M. Let us fix $s > \frac{1}{4}$ and*

$$(u_1, u_2) = \sum_n (u_{n,1} e_n(x), u_{n,2} e_n(x)) \in H^s(M) \times H^{s-1}(M).$$

Let (g_n) and (h_n) be two families or independent random variables satisfying the assumptions in Theorem 2.1. Consider

$$(u_0^\omega, u_1^\omega) = \sum_n (u_{n,1} g_n^\omega e_n(x), h_n^\omega u_{n,2} e_n(x))$$

the associated random function. Then for almost every initial data (u_0^ω, u_1^ω), there exists $T > 0$ such that there exists a unique solution u of (18) in a space continuously embedded in $C([-T, T]; H^s(M))$, and furthermore, there exist $C > 0, \delta > 0$ such that

$$p(T \geq T_0) \geq 1 - Ce^{-c/T_0^\delta}, \quad c, \delta > 0. \tag{19}$$

Remark. *Notice that if the initial data (u_0, u_1) are in $\mathcal{H}^s(M) \times \mathcal{H}^{s-1}(M)$ but* not *in $\mathcal{H}^\sigma(M) \times \mathcal{H}^{\sigma-1}(M)$, then almost surely (u_0^ω, u_1^ω) are* not *in $\mathcal{H}^\sigma(M) \times \mathcal{H}^{\sigma-1}(M)$. Consequently, this theorem provides us with a large number of initial data of super-critical regularity, for which local existence of a strong solution holds.*

Sketch of proof. Let us recall first how, using purely deterministic arguments, one can prove that (18) is locally well posed for initial data in $\mathcal{H}^s(M) \times \mathcal{H}^{s-1}(M)$ when $s \geq 1/2$. We shall use the following Strichartz estimate due to Kapitanskii [26]

Theorem 3.2. *The solution of the linear wave equation*

$$(\partial_t^2 - \Delta) u = f, \quad (u, \partial_t u)|_{t=0} = (u_1, u_2) \tag{20}$$

satisfies

$$\|u\|_{L^4((0,T\times\Omega))} \leq C\Big(\|(u_1, u_2)\|_{H^{1/2}(M)\times H^{-1/2}(M)} + \|f\|_{L^{4/3}((0,T)\times M)} \Big).$$

Now, to solve (18), we simply look for a fixed point of the operator

$$K : u \mapsto \cos(t\sqrt{-\boldsymbol{\Delta}}) u_1 + \frac{\sin(t\sqrt{-\boldsymbol{\Delta}})}{\sqrt{-\boldsymbol{\Delta}}} u_2 + \int_0^t \frac{\sin((t-s)\sqrt{-\boldsymbol{\Delta}})}{\sqrt{-\boldsymbol{\Delta}}} u^3(s) ds$$

in the space $C^0((0,T);H^s(M)) \cap L^4((0,T)\times M)$, and using Theorem 3.2, it is not difficult to see the existence of such a fixed point (notice that $u \in L^4 \Rightarrow u^3 \in L^{4/3}$). The idea of the proof of Theorem 3.1 is now very simple. Instead of performing, the previous fixed point in the Strichartz type space, we perform a first iteration and search for a solution under the form

$$u = \cos(t\sqrt{-\mathbf{\Delta}})u_1 + \frac{\sin(t\sqrt{-\mathbf{\Delta}})}{\sqrt{-\mathbf{\Delta}}}u_2 + v = u_{\text{free}} + v.$$

The function v is solution of

$$(\partial_t^2 - \mathbf{\Delta})v + (u_{\text{free}} + v)^3 = 0, \quad (v, \partial_t v)|_{t=0} = (0,0) \tag{21}$$

and we can rewrite this equation as a fixed point

$$v = \widetilde{K}(v) = \int_0^t \frac{\sin((t-s)\sqrt{-\mathbf{\Delta}})}{\sqrt{-\mathbf{\Delta}}}(u_{\text{free}} + v)^3(s)ds.$$

Now, according to Theorem 2.5, almost surely, there exists $T > 0$ such that

$$u_{\text{free}} \in L^4((0,T)\times M)$$

(notice that the additional time dependence plays no role and the proof of Theorem 2.1 applies). As a consequence, the same proof as in the deterministic case for the operator K applies and shows the existence of a fixed point for $\widetilde{K}$.

3.2. A global existence result. Having the previous local result in mind, a natural question is whether one can exhibit cases where it is possible to prove global (in time) existence for the solutions. It turns out that it is the case for a very particular model problem: Consider the case of the wave equation in the unit ball of $\mathbb{R}^3$, $\mathbf{B}$, with Dirichlet boundary conditions

$$(\partial_t^2 - \Delta)u + |u|^{p-1}u = 0, \quad u\,|_{\partial\mathbf{B}} = 0, \qquad (u, \partial_t u)|_{t=0} = (u_1, u_2) \tag{22}$$

In this case, the critical index is

$$s_c = \frac{3}{2} - \frac{2}{p-1},$$

and for $k > 3$, $s_c > \frac{1}{2}$. Consider now $(e_n)_{n\in\mathbb{N}}$ the sequence of *radial* eigenfunctions of the Laplace operator with Dirichlet boundary conditions in $\mathbf{B}$. The following result [14] shows that the Cauchy problem is, in this particular case globally well posed for a large number of super-critical initial data.

Theorem 3.3. *Suppose that $k < 4$. Fix a real number p such that $\max(4, 2\alpha) < p < 6$. Let $((h_n(\omega), l_n(\omega))_{n=1}^{\infty}$ be a sequence of independent standard real Gaus-*

sian random variables on a probability space $(\Omega, \mathcal{A}, p)$*. Consider (22) with initial data*

$$f_1^\omega(r) = \sum_{n=1}^{\infty} \frac{h_n(\omega)}{n\pi} e_n(r), \quad f_2^\omega(r) = \sum_{n=1}^{\infty} l_n(\omega) e_n(r)\,, \tag{23}$$

where $(e_n(r))_{n=1}^{\infty}$ *is the orthonormal basis consisting in radial eigenfunctions of the Laplace operator with Dirichlet boundary conditions, associated to eigenvalues* $-(\pi n)^2$*. Then for every* $s < 1/2$*, almost surely in* $\omega \in \Omega$*, the problem (22) has a unique global solution*

$$u^\omega \in C(\mathbb{R}, H^s(\mathbf{B})) \cap L^p_{loc}(\mathbb{R}_t; L^p(\mathbf{B}))\,.$$

Furthermore, the solution is a perturbation of the linear *solution*

$$u^\omega(t) = U(t)(f_1^\omega, f_2^\omega) + v^\omega(t) = \cos(t\sqrt{-\mathbf{\Delta}})f_1^\omega + \frac{\sin(t\sqrt{-\mathbf{\Delta}})}{\sqrt{-\mathbf{\Delta}}} f_2^\omega + v^\omega(t),$$

where $v^\omega \in C(\mathbb{R}, H^\sigma(\mathbf{B}))$ *for some* $\sigma > 1/2$*. Moreover there exists* $C > 0$*, and almost surely* D^ω *such that*

$$\|u^\omega(t)\|_{H^s(\mathbf{B})} \leq C \log(D^\omega + |t|)^{\frac{1}{2}}, \mathcal{P}(D^\omega > \Lambda) \leq C e^{-\lambda^2/C}\,.$$

Notice that as soon as $s < 1/2$, the initial data given by (23) are almost surely in $H^s(\mathbf{B}) \times H^{s-1}(\mathbf{B})$ and as soon as $s \geq \frac{1}{2}$ are almost surely not in $H^s(\mathbf{B}) \times H^{s-1}(\mathbf{B})$, and consequently in the range of non linearities $3 < k < 4$, the initial data we consider are super-critical. The proof of this result combines a local Cauchy at the probabilistic level with the Gibbs measure strategy performed by Bourgain [7], following the trend iniciated by Lebowitz-Rose-Speer [34].

4. Non Linear Harmonic Oscillators

In this section, I will present some results on the non linear harmonic oscillator

$$\begin{cases} i\partial_t u + \partial_x^2 u - x^2 u = \kappa_0 |u|^{k-1} u, \quad (t,x) \in \mathbb{R} \times \mathbb{R}, \\ u(0,x) = f(x), \end{cases} \tag{24}$$

where $k \geq 3$ is an odd integer and where either $\kappa_0 = 1$ or $\kappa_0 = -1$. Our main result [12] shows once again that the Cauchy problem is globally well posed for a large number of initial data.

Theorem 4.1. *Consider the* L^2 *Wiener measure on* $\mathcal{D}'(\mathbb{R})$*,* μ*, constructed on the harmonic oscillator eigenbasis, i.e.* μ *is the distribution of the random*

variable

$$\sum_{n=0}^{\infty} \sqrt{\frac{2}{2n+1}} g_n(\omega) h_n(x),$$

where $(h_n)_{n=0}^{\infty}$ are the Hermite functions and $(g_n)_{n=0}^{\infty}$ is a system of standard independent complex Gaussian random variables. Then in the defocusing case, for any order of nonlinearity $k < +\infty$, and in the focusing case for the cubic non linearity, the Cauchy problem (24) is globally well posed for μ-almost every initial data. Furthermore, in both cases, there exists a Gibbs measure, absolutely continuous with respect to μ, which is invariant by this flow.

An interesting by-product of our analysis is the following result for the L^2 critical and super-critical equation

$$\begin{cases} i\partial_t u + \partial_x^2 u = |u|^{k-1}u, \quad k \geq 5, \quad (t,x) \in \mathbb{R} \times \mathbb{R}, \\ u(0,x) = u_0(x) \end{cases} \tag{25}$$

Theorem 4.2. *[[12]] For any $0 < s < 1/2$, the equation (25) has for μ-almost every initial data a unique global solution satisfying*

$$u(t,\cdot) - e^{-it\mathbf{\Delta}} u_0 \in C\big(\mathbb{R}; \mathcal{H}^s(\mathbb{R})\big)$$

(the uniqueness holds in a space continuously embedded in $C\big(\mathbb{R}; \mathcal{H}^s(\mathbb{R})\big)$). Moreover, the solution scatters in the following sense. There exists μ-almost surely states $g_\pm \in \mathcal{H}^s(\mathbb{R})$ so that

$$\|u(t,\cdot) - e^{it\mathbf{\Delta}}(f + g_\pm)\|_{\mathcal{H}^s(\mathbb{R})} \longrightarrow 0, \quad \textit{when} \quad t \longrightarrow \pm\infty.$$

Remark. *Notice that Theorem 4.2 gives global existence whilst no invariant measure is involved in the proof (see Colliander-Oh [23, 22] for results in this direction).*

The proof of Theorem 4.2 uses the pseudo-conformal transform (see [15] for a use of this transform in the context of L^2 scattering problems).

Proposition 4.3. *Suppose that $v(s,y)$ is a solution of the problem*

$$i\partial_s v + \partial_y^2 v = |v|^{k-1}v, \quad s \in \mathbb{R}, \quad y \in \mathbb{R}. \tag{26}$$

We define $u(t,x) = \mathcal{L}(v)(t,x)$ for $|t| < \frac{\pi}{4}$, $x \in \mathbb{R}$ by

$$u(t,x) = \frac{1}{\cos^{\frac{1}{2}}(2t)} v\left(\frac{\tan(2t)}{2}, \frac{x}{\cos(2t)}\right) e^{-\frac{ix^2 \mathrm{tg}(2t)}{2}}. \tag{27}$$

Then u solves the problem

$$i\partial_t u - (\partial_x^2 - x^2)u = \cos^{\frac{k-5}{2}}(2t)|u|^{k-1}u, \quad |t| < \frac{\pi}{4},\, x \in \mathbb{R}. \tag{28}$$

As a consequence, in the case $k = 5$, (28) reduces to (25), and Theorem 4.2 follows rather directly from Theorem 4.1 In the case $k \geq 7$, the proof is more involved and relies on an analog of Theorem 4.1 for (28) (notice that this latter equation is non autonomous).

Sketch of proof of Theorem 4.1. For low order non linearities ($p \leq 7$), the proof follows the same lines as in the case of wave equations, and relies on Theorem 2.6 (or more precisely on similar estimates for the solution of the linear harmonic Schrödinger equation $u = e^{itH}u_0$). However, as soon as $p \geq 9$, these estimates are not sufficient, because they allow only for a gain of at most $1/4$ space derivatives, and the exponent for which $s_c = \frac{1}{4}$ is precisely $k = 9$. As a consequence, our analysis requires a full bi-linear analysis at the probabilistic level. The bilinear nature of our probabilistic analysis can be seen though the following statement which shows that by considering nonlinear quantities, a gain of (almost) $1/2$ space derivatives can be achieved.

$$\forall\, \theta < 1/2,\ \forall\, t \in \mathbb{R}, \quad \|(e^{-itH}u^\omega)^2\|_{\mathcal{H}^\theta} < +\infty, \quad \mu \text{ almost surely.} \tag{29}$$

4.1. Bilinear estimates. In this section we give a proof of (29) which was pointed to us by P. Gérard. Observe that (29), applied with $t = 0$ implies that $(u_0^\omega)^2(x)$ is a.s. in $\mathcal{H}^\theta$ for every $\theta < 1/2$ which is a remarkable smoothing property satisfied by the random series $(u_0^\omega)(x)$. The key point in the proof of (29) is the following bilinear estimate for Hermite functions.

Lemma 4.4. *There exists $C > 0$ so that for all $0 \leq \theta \leq 1$ and $n, m \in \mathbb{N}$*

$$\|h_n\, h_m\|_{\mathcal{H}^\theta(\mathbb{R})} \leq C \max(n,m)^{-\frac{1}{4}+\frac{\theta}{2}} \big(\log\big(\min(n,m)+1\big)\big)^{\frac{1}{2}}. \tag{30}$$

Proof. We give an argument we learned from Patrick Gérard. It suffices to prove (30) for $\theta = 0$ and $\theta = 1$ (the general case then follows by interpolation). The case $\theta = 1$ can actually be directly reduced to the case $\theta = 0$ by taking space derivatives. We are going to use the generating series:

$$\begin{aligned} E(x,y,\alpha) &= \sum_{n\geq 0} \alpha^n\, h_n(x)\, h_n(y) \\ &= \frac{1}{\sqrt{\pi(1-\alpha^2)}}\, \exp\, \left(-\frac{1-\alpha}{1+\alpha}\, \frac{(x+y)^2}{4} - \frac{1+\alpha}{1-\alpha}\, \frac{(x-y)^2}{4} \right). \end{aligned} \tag{31}$$

Therefore, if we set

$$I(\alpha,\beta) \equiv \int_{\mathbb{R}} E(x,x,\alpha)E(x,x,\beta)\mathrm{d}x,$$

then we get

$$I(\alpha,\beta)=\frac{1}{\pi}(1-\alpha^2)^{-\frac12}(1-\beta^2)^{-\frac12}\int_{\mathbb{R}} e^{-\left(\frac{1-\alpha}{1+\alpha}+\frac{1-\beta}{1+\beta}\right)x^2}\mathrm{d}x$$
$$=\frac{1}{\sqrt{2\pi}}(1-\alpha)^{-\frac12}(1-\beta)^{-\frac12}(1-\alpha\beta)^{-\frac12}. \quad (32)$$

On the other hand, coming back to the definition

$$I(\alpha,\beta)=\sum_{n,m\geq 0}\alpha^n\beta^m\int_{\mathbb{R}} h_n^2(x)\,h_m^2(x)\mathrm{d}x.$$

Hence to get a useful expression for the L^2 norm of the product of two Hermite functions, it suffices to expand (32) in entire series in α and β. Write

$$(1-x)^{-\frac12}=\sum_{p\geq 0}c_p x^p,\quad c_0=1,\qquad c_p=\frac{(2p-1)\,!}{2^{2p-1}\,p\,!\,(p-1)\,!},\quad p\geq 1.$$

Therefore, by the Stirling formula, there exists $C>0$ so that $|c_p|\leq\frac{C}{\sqrt{p+1}}$ for all $p\geq 0$. Now by (32) and the previous estimate

$$\int_{\mathbb{R}} h_n^2(x)\,h_m^2(x)\mathrm{d}x=\frac{1}{\sqrt{2\pi}}\sum_{\substack{p,q,r\geq 0\\ p+r=n,\ q+r=m}} c_p\,c_q\,c_r$$
$$\leq C\sum_{0\leq r\leq\min(n,m)}(n-r+1)^{-\frac12}\,(m-r+1)^{-\frac12}\,(r+1)^{-\frac12}.$$

Without restricting the generality we may suppose that $m\geq n$. If $m\leq 2n$ then we obtain the needed bound by considering separately the cases when the sum runs over $r<m/2$ and $r\geq m/2$. If $m>2n$, then we can write $(m-r+1)^{-\frac12}\leq c(1+m)^{-\frac12}$ and the needed bound follows directly. Therefore we get (30) in the case $\theta=0$. This completes the proof of Lemma 4.4. □

Denote by $u^\omega_{\mathrm{free}}(t,x)$ the free Schrödinger solution with initial condition $u_0^\omega\phi(\omega,x)$, i.e.

$$u^\omega_{\mathrm{free}}(t,x)=e^{-itH}u_0^\omega=\sum_{n\geq 0}\frac{\sqrt{2}}{\lambda_n}e^{-it\lambda_n^2}\,g_n^\omega h_n(x).$$

Write the decomposition $u=\sum_N u_N$, where the summation is taken over the dyadic integers and for N a dyadic integer

$$u_N(\omega,t,x)=\sum_{N\leq n<2N}\alpha_n(t)h_n(x)g_n^\omega,\quad \alpha_n(t)=\sqrt{\frac{2}{2n+1}}e^{-i(2n+1)t}\,.$$

Let us fix $t \in \mathbb{R}$ and $0 \leq \theta < \frac{1}{2}$. It suffices to show that the expression

$$J(t,x,\omega) \equiv \left| \sum_M \sum_N H^{\theta/2}\,(u_N\, u_M) \right|$$

belongs to $L^2(\mathbb{R} \times \Omega)$ (here the summation is again taken over the dyadic values of M, N). Using the Cauchy-Schwarz inequality, a symmetry argument and summing geometric series, for all $\epsilon > 0$ we can write

$$J(t,x,\omega) \leq C \left(\sum_{N \leq M} M^{\epsilon} |H^{\theta/2}\,(u_N\, u_M)\,|^2 \right)^{\frac{1}{2}}. \tag{33}$$

Coming back to the definition we can write

$$H^{\theta/2}\big(u_N\, u_M\big) = \sum_{\substack{N \leq n \leq 2N \\ M \leq m \leq 2M}} \alpha_n\, \alpha_m\, g_n\, g_m\, H^{\theta/2}\big(h_n\, h_m\big).$$

We now estimate $\mathbb{E}(|H^{\theta/2}\big(u_N\, u_M\big)|^2$. We make the expansion

$$|H^{\theta/2}\big(u_N\, u_M\big)|^2 = \sum_{\substack{N \leq n_1, n_2 \leq 2N \\ M \leq m_1, m_2 \leq 2M}} \alpha_{n_1}\, \overline{\alpha}_{n_2}\, \alpha_{m_1}\, \overline{\alpha_{m_2}}\, g_{n_1}\, \overline{g_{n_2}}\, g_{m_1}\, \overline{g_{m_2}}\, H^{\theta/2}\big(h_{n_1}\, h_{m_1}\big)\, \overline{H^{\theta/2}\big(h_{n_2}\, h_{m_2}\big)}.$$

The random variables g_n are centered and independent, and consequently, we have $\mathbb{E}\big[\, g_{n_1}\, \overline{g_{n_2}}\, g_{m_1}\, \overline{g_{m_2}}\,\big] = 0$, unless the indexes are pairwise equal (i.e. ($n_1 = n_2$ and $m_1 = m_2$), or ($n_1 = m_2$ and $n_2 = m_1$). This implies that

$$\mathbb{E}(|H^{\theta/2}\big(u_N\, u_M\big)|^2) \leq C \sum_{\substack{N \leq n \leq 2N \\ M \leq m \leq 2M}} |\alpha_n|^2 |\alpha_m|^2 |H^{\theta/2}\big(h_n\, h_m\big)|^2. \tag{34}$$

We integrate (34) in x and by (30) we deduce that for all $\epsilon > 0$

$$\begin{aligned} \int_{\Omega \times \mathbb{R}} |H^{\theta/2}(u_N\, u_M)|^2 &\leq C \sum_{\substack{N \leq n \leq 2N \\ M \leq m \leq 2M}} |\alpha_n|^2 |\alpha_m|^2 \int_{\mathbb{R}} |H^{\theta/2}\big(h_n\, h_m\big)|^2 \mathrm{d}x \\ &\leq C \sum_{\substack{N \leq n \leq 2N \\ M \leq m \leq 2M}} (\max\,(M,N))^{-\frac{1}{2}+\theta+\epsilon} |\alpha_n|^2 |\alpha_m|^2. \end{aligned}$$

Therefore using that $|\alpha_n| \leq \langle n\rangle^{-\frac{1}{2}}$, we get

$$\begin{aligned}\mathbb{E}(J(t,x,\omega)^2) &\leq C\sum_{N\leq M}\sum_{\substack{N\leq n\leq 2N\\ M\leq m\leq 2M}} M^{-\frac{1}{2}+\theta+2\epsilon}|\alpha_n|^2|\alpha_m|^2\\ &\leq C\sum_{N\leq M}\sum_{\substack{N\leq n\leq 2N\\ M\leq m\leq 2M}} M^{-\frac{1}{2}+\theta+2\epsilon}(MN)^{-1}<\infty,\end{aligned}$$

provided ϵ is small enough, namely ϵ such that $-\frac{1}{2}+\theta+2\epsilon<0$. This completes the proof of (29).

5. Improved Sobolev Embeddings

As shown in the previous section, our applications to partial differential equations of Paley-Zygmund's result rely on the simple observation that "typical" functions n $\mathcal{H}^s(M)$ enjoy better L^p properties than what the Sobolev embeddings would predict. Namely, the L^∞ norm is essentially bounded (modulo logarithmic loss) by the $\mathcal{H}^{\frac{d-1}{2}}$ norm (versus the $\mathcal{H}^{d/2}$ norm for classical Sobolev embeddings). Notice that this bound is improved, in the case of the tora $\mathbb{T}^d$ all the way down to $\mathcal{H}^0$. In this section, I will present some other randomizations obtained with G. Lebeau [11] on compact manifolds. Notice that other applications to linear and non linear problems are developped in [11], and we expect these constructions to be of interest in view of further applications to partial differential equations.

5.1. Construction of the measure. Let M be a compact riemanian manifold, let $I=[c,c']$, $0\leq c<c'<\infty$ be an interval and $E_{I,h}$ the subspace of $L^2(M)$ of dimension $N(I,h)$ defined by

$$E_{I,h}=\left\{u=\sum_{k\in I_h} z_k e_k(x),\ z_k\in\mathbb{C}\right\},\quad I_h=\{k,\ h\omega_k\in I\}. \tag{35}$$

According to the precised Weyl formula (see [21, Theorem 1.1]), we have for $h\in]0,1]$

$$N(I,h)=(2\pi h)^{-d}\mathrm{Vol}(M)\mathrm{Vol}(S^{d-1})\int_I \rho^{d-1}d\rho+O(h^{-d+1}). \tag{36}$$

Let us recall that Sobolev injections read

$$\|u\|_{L^\infty(M)}\leq Ch^{-d/2}\|u\|_{L^2(M)}\quad \forall u\in E_{I,h}. \tag{37}$$

Notice that these estimates are optimal as can easily bee seen by considering

the sequence $h^{-d/2}\chi(x/h)$, where $\chi \in C_0^\infty$ a fixed function (in a local coordinate chart). Denote by $S_{I,h}$ the unit sphere of the euclidean space $E_{I,h} = \mathbb{C}^{N(I,h)}$, and $P_{I,h}$ the uniform probability on $S_{I,h}$. We can now define probability measures on $E_{I,h}$ by picking a probability measure in the radial variable $\rho(r)$, with sufficient fast decay near infinity (e.g. Gaussian), and defining

$$d\mu_{I,h} = dP_{I,h} \otimes d\rho.$$

A typical example (to which all other examples reduce eventually) is of course the simplest choice $\rho = \delta_{r=1}$ for which the measure $\mu_{I,h}$ is simply the uniform measure on the unit sphere of $E_{I,h}$, which in the sequel will still be denoted by $P_{I,h}$. Finally, taking any family of positive real numbers $(\alpha_n) > 0$, we can rescale (in the radial variable) the measure by defining

$$d\mu_{I,h,\alpha_h} = dP_{I,h} \otimes \alpha_h d\rho(\frac{r}{\alpha_h}).$$

The choice $I = [1/2, 2[, h_k = 2^{-k}$, $k \in \mathbb{N}$ (with a suitable modification for $k = 0$) gives

$$L^2(M) = \left\{ u = \sum_k u_k; u_k \in E_{I,h_k}; \sum_k \|u_k\|_{L^2}^2 < +\infty \right\}$$

and the Sobolev space $H^s(M)$ can also be expressed in terms of this decomposition

$$H^s(M) = \left\{ u = \sum_k u_k; u_k \in E_{I,h_k}; \sum_k 2^{2ks} \|u_k\|_{L^2}^2 < +\infty \right\}.$$

As a consequence, the choice of $\alpha_{h_k} = 2^{-k}\beta_k$ with $\beta_k \in \ell^2$ ensures that the measure

$$d\mu_{s,(\beta_n)} = \otimes_k d\mu_{I,h_k,\alpha_{h_k}}$$

defines a measure on $\oplus_k E_{I,h_k}$ which is supported by $H^s(M)$.

5.2. Improved Sobolev embeddings.

The measures constructed in the previous section satisfy:

Theorem 5.1.

- *For any choice of sequence $(\beta_n) \in \ell^2$, the measure $d\mu_{s,(\beta_n)}$ is supported in $H^s(M)$.*
- *For any $s' > s$, if the sequence (β_n) satisfies*

$$\sum_n |\beta_n|^2 (1 + 2^{2(s'-s)}) = +\infty,$$

then the space $H^{s'}(M)$ has $d\mu_{s,(\beta_n)}$-measure equal to 0.

- *For any $s > 0$, the measure $d\mu_{s,(\beta_n)}$ is supported in $L^\infty(M)$. In other words, "for any $\epsilon > 0$, almost surely, $H^\epsilon(M) \subset L^\infty(M)$".*

Remark. *A similar result was obtained by Shiffman-Zelditch in [44] in the different context of random sequences of holomorphic sections of high powers of a positive line bundle.*

The main step for the proof of Theorem 5.1 is the proof of the following semi-classical result

Theorem 5.2. *For any $c < Vol(M)$, there exists $C > 0$ such that for any $h \in (0, 1]$, and any $\lambda > 0$,*

$$P_{I,h}(\{u \in E_{I,h}; \|u\|_{L^\infty} > \lambda\}) \leq Ch^{-d(1+d/2)} e^{-c_2\lambda^2} \tag{38}$$

Indeed, taking $\lambda = h^{-\epsilon}$ in (38), we obtain

$$P_{I,h}(\{u \in E_{I,h}; \|u\|_{L^\infty} > h^{-\epsilon}\}) \leq C' e^{-c' h^{-2\epsilon}},$$

and Theorem 5.1 follows after suitable resummations. Now, in turn, Theorem 5.2 follows from the classical concentration of measure phenomenon (see Ledoux [35])

Theorem 5.3. *Consider a Lipshitz function F, on the N dimensional sphere $\mathbb{S}^N$, endowed with its natural geodesic metric, and with the uniform probability measure μ_N. Let us define the mediane, $\mathcal{M}(F)$, of the function F by the relation*

$$\mu_N(\{x \in \mathbb{S}^N; F(x) \geq \mathcal{M}(F)\}) \geq \frac{1}{2}, \qquad \mu(\{x \in \mathbb{S}^N; F(x) \leq \mathcal{M}(F)\}) \geq \frac{1}{2}.$$

Then for any $r > 0$,

$$\mu(\{x \in \mathbb{S}^N;\ F(x) - \mathcal{M}(F)| > r\}) \leq 2e^{-(N-1)\frac{r^2}{\|f\|^2_{Lip}}}.$$

References

[1] T. Alazard and R. Carles. Loss of regularity for supercritical nonlinear Schrödinger equations. *Math. Ann.*, 343(2):397–420, 2009.

[2] A. Ayache and N. Tzvetkov. L^p properties for Gaussian random series. *Trans. Amer. Math. Soc.*, 360(8):4425–4439, 2008.

[3] J. Bourgain. Exponential sums and nonlinear Schrödinger equations. *Geom. and Funct. Anal.*, 3:157–178, 1993.

[4] J. Bourgain. Fourier transform restriction phenomena for certain lattice subsets and application to nonlinear evolution equations i. Schrödinger equations. *GAFA*, 3:107–156, 1993.

[5] J. Bourgain. Global solutions of nonlinear Schrödinger equations. *Colloq. Publications, American Math. Soc.*, 1999.

[6] J. Bourgain. Global wellposedness of defocusing critical nonlinear Schrödinger equations in the radial case. *J. Amer. Math. Soc.*, 12:145–171, 1999.

[7] J. Bourgain. Invariant measures for the 2D-defocusing nonlinear Schrödinger equation. *Comm. Math. Phys.*, 176(2):421–445, 1996.

[8] N. Burq, P. Gérard, and N. Tzvetkov. Bilinear eigenfunction estimates and the nonlinear Schrödinger equation on surfaces. *Inventiones Mathematicae*, 159(1):187 – 223, 2005.

[9] N. Burq, P. Gérard, and N. Tzvetkov. Multilinear eigenfunction estimates and global existence for the three dimensional nonlinear Schrödinger equations. *Ann. Sci. École Norm. Sup. (4)*, 38(2):255–301, 2005.

[10] N. Burq, Patrick Gérard, and N. Tzvetkov. Global solutions for the nonlinear Schrödinger equation on three-dimensional compact manifolds. *Mathematical aspects of nonlinear PDE's. Annals of Mathematical Studies*, 2006. to appear.

[11] N. Burq and G. Lebeau. Injections de sobolev probabilistes et applications. *in preparation*, 20010.

[12] N. Burq, L. Thomann, and N. Tzvetkov. Long time dynamics for one dimensional non linear schrödinger equations. *preprint, http://fr.arxiv.org/abs/1002.4054*.

[13] N. Burq and N. Tzvetkov. Random data cauchy theory for supercritical wave equations I: Local theory. *Inventiones Mathematicae*, 173:449–475, 2008.

[14] N. Burq and N. Tzvetkov. Random data cauchy theory for supercritical wave equations II: A global result. *Inventiones Mathematicae*, 173:477–496, 2008.

[15] R. Carles. Remarks on nonlinear Schrödinger equations with harmonic potential. *Ann. Henri Poincaré*, 3(4):757–772, 2002.

[16] M. Christ, J. Colliander, and T. Tao. Asymptotics, frequency modulation, and low regularity ill-posedness for canonical defocusing equations. *Amer. J. Math.*, 125(6):1235–1293, 2003.

[17] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao. Global well-posedness and scattering for the energy-critical nonlinear Schrödinger equation in $\mathbb{R}^3$. *Ann. of Math. (2)*, 167(3):767–865, 2008.

[18] J. Dziubański and P. Głowacki. Sobolev spaces related to Schrödinger operators with polynomial potentials. *Math. Z.*, 262(4):881–894, 2009.

[19] P. Gérard. Nonlinear Schrödinger equations in inhomogeneous media: wellposedness and illposedness of the Cauchy problem. In *International Congress of Mathematicians. Vol. III*, pages 157–182. Eur. Math. Soc., Zürich, 2006.

[20] J. Ginibre and G. Velo. The global Cauchy problem for the nonlinear Schrödinger equation. *Ann. I.H.P. (Anal. non lin.)*, 2:309–327, 1985.

[21] L. Hörmander. The spectral function of an elliptic operator. *Acta Math.*, 121:193–218, 1968.

[22] J. Colliander and T. Oh. Almost sure global solutions of the periodic cubic nonlinear schrödinger equation below l2,. preprint.

[23] J.Colliander and T. Oh. Almost sure local well-posedness of the periodic cubic nonlinear schrödinger equation below l2. preprint.

[24] J.-P. Kahane. *Some random series of functions*, volume 5 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 1985.

[25] L. Kapitanskii. Some generalizations of the Strichartz-Brenner inequality. *Leningrad Math. J.*, pages 693–726, 1990.

[26] L. V. Kapitanskiĭ. Some generalizations of the Strichartz-Brenner inequality. *Algebra i Analiz*, 1(3):127–159, 1989.

[27] A. Kolmogorov and G. Seliverstoff. Sur la convergence des séries de Fourier *Comptes rendus* CLXXVIII, 303–306, 1924.

[28] M. Keel and T. Tao. Endpoint Strichartz estimates. *Amer. Jour. of Math.*, pages 955–980, 1998.

[29] C. E. Kenig and F. Merle. Global well-posedness, scattering and blow-up for the energy-critical, focusing, non-linear Schrödinger equation in the radial case. *Invent. Math.*, 166(3):645–675, 2006.

[30] C. E. Kenig and F. Merle. Global well-posedness, scattering and blow-up for the energy-critical focusing non-linear wave equation. *Acta Math.*, 201(2):147–212, 2008.

[31] H. Koch and D. Tataru. L^p eigenfunction bounds for the Hermite operator. *Duke Math. J.*, 128(2):369–392, 2005.

[32] G. Lebeau. Non linear optic and supercritical wave equation. Hommage Pascal Laubin, *Bull. Soc. Roy. Sci. Lige*, 70 (4–6), 267–306, 2002.

[33] G. Lebeau. Perte de rgularit pour les quations d'ondes sur-critiques, *Bull. Soc. Math. France*, 133 (1), 145–157, 2005

[34] J. Lebowitz, R. Rose, E. Speer. Statistical dynamics of the nonlinear Schrödinger equation. *J. Stat. Physics*, V 50 (1988) 657–687.

[35] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[36] M. B. Marcus and G. Pisier. *Random Fourier series with applications to harmonic analysis*, volume 101 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, N.J., 1981.

[37] F. Merle and P. Raphael. Sharp upper bound on the blow-up rate for the critical nonlinear Schrödinger equation. *Geom. Funct. Anal.*, 13(3):591–642, 2003.

[38] F. Merle and P. Raphael. On universality of blow-up profile for L^2 critical nonlinear Schrödinger equation. *Invent. Math.*, 156(3):565–672, 2004.

[39] F. Merle and P. Raphael. The blow-up dynamic and upper bound on the blow-up rate for critical nonlinear Schrödinger equation. *Ann. of Math. (2)*, 161(1):157–222, 2005.

[40] F. Merle and P. Raphael. Profiles and quantization of the blow up mass for critical nonlinear Schrödinger equation. *Comm. Math. Phys.*, 253(3):675–704, 2005.

[41] Y.-G. Oh. Cauchy problem and Ehrenfest's law of nonlinear Schrödinger equations with potentials. *J. Differential Equations*, 81(2):255–274, 1989.

[42] R.E.A.C. Paley and A. Zygmund. On some series of functions (1) (2) (3). *Proc. Camb. Phil. Soc.*, (26–28):337–357, 458–474, 190–205, 1930–1932.

[43] H. Rademacher Einige Satze über Reihen yon allgemeinen Orthogonal. funktionen. *Math. Annalen* 87, S:112–138, 1922.

[44] B. Shiffman and S. Zelditch. Random polynomials of high degree and Levy concentration of measure. *Asian J. Math.*, 7(4):627–646, 2003.

[45] C. Sogge. Concerning the L^p norm of spectral clusters for second order elliptic operators on compact manifolds. *Jour. of Funct. Anal.*, 77:123–138, 1988.

[46] R. S. Strichartz. Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations. *Duke Math. J.*, 44(3):705–714, 1977.

[47] N. Tzvetkov Invariant measures for the defocusing nonlinear Schrödinger equation. *Ann. Inst. Fourier*, 58, 7: 2543–2604, 2008.

[48] L. Thomann. Instabilities for supercritical Schrödinger equations in analytic manifolds. *J. Differential Equations*, 245(1):249–280, 2008.

[49] K. Yajima and G. Zhang. Strichartz inequality and smoothing property for Schrödinger equations with potential superquadratic at infinity. *Sūrikaisekikenkyūsho Kōkyūroku*, (1234):179–194, 2001. Tosio Kato's method and principle for evolution equations in mathematical physics (Sapporo, 2001).

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Study of Multidimensional Systems of Conservation Laws: Problems, Difficulties and Progress

Shuxing Chen*

Abstract

In the study of multidimensional systems of conservation laws people confront more difficulties than that for one-dimensional systems. The difficulties include characteristic boundary, free boundary associated with unknown nonlinear waves, various nonlinear wave structure, mixed type equations, strong singularities, etc. Most of them come from the complexity of characteristics. We will give a survey on the progress obtained in the study of this topic with the applications in various physical problems, and will also emphasize some crucial points for the further development of this theory in future.

Mathematics Subject Classification (2010). Primary 35L65; Secondary 35L67; 35L60; 76N15; 35M10.

Keywords. Conservation laws; characteristics; free boundary value problem; shock; transonic flow; mixed type equation.

1. Introduction

The system of conservation laws originates from the study of compressible fluid dynamics, shallow water waves, elastrodynamics, magnetohydrodynamics etc. The extensive study of it started from the middle of the last century. The problems in its first studies are one-space-dimensional. Since the real physical problems occur in the three-dimensional space, then in the study of one-space-dimensional problems it is automatically assumed that all quantities under consideration is uniform with respect to two space-variables among three space-variables in regular physical space. However, many physical problems do not

*School of Mathematical Sciences, Fudan University, Shanghai, 200433, People's Republic of China. E-mail: sxchen@fudan.edu.cn

have such symmetry, so that they are generally multidimensional. Therefore, the study of multidimensional conservation laws is inevitable and is indisputably important.

In the study of multidimensional systems of conservation laws people often confronts more difficulties than that for one-dimensional systems. Most of difficulties come from the complicated structure of their characteristics. As a matter of fact, the characteristics are the path of the propagation of perturbations in the motion of media. In one-space-dimensional case the characteristics of the system under consideration are characteristic curves in time-space plane, but in multidimensional case the characteristic variaties are surfaces or more general manifolds, so that the location of characteristics or their intersections will be rather complicated. This situation is also reflected in the descriptions of various nonlinear waves associated with characteristics, including shocks, rarefaction waves and contact discontinuities. As for the methodology, the integration along characteristic curves is a very efficient method in treating various one-space-dimensional problems. Based on it many techniques are developed. However, due to the complexity of characteristics one needs to develop totally different methods to deal with corresponding problems in multidimensional case. Next we will recall the progress obtained in the study of multidimensional conservation laws with showing the main difficulties, which have been overcome or to be overcome.

In this paper the main prototype of the system under consideration is the Euler system for inviscid compressible flow with the form

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \ div(\rho \vec{v}) = 0, \\ \dfrac{\partial (\rho \vec{v})}{\partial t} + \ div(\rho \vec{v} \times \vec{v}) + \nabla p = 0, \\ \dfrac{\partial E}{\partial t} + div(\vec{v}(E+p)) = 0, \end{cases} \tag{1}$$

where $\rho, \vec{v}, p, E$ represent the density, the velocity, the pressure and the total energy respectively. Meanwhile, when the flow is isentropic and irrotational, the Euler system can be reduced to a second order equation – potential flow equation (see [34])

$$(\rho(\nabla \phi))_t + \sum_{j=1}^{3} (\phi_{x_j} \rho(\nabla \phi))_{x_j} = 0, \tag{2}$$

where ϕ is the potential of velocity, satisfying $\vec{v} = \nabla \phi$.

2. Characteristic Boundary Value Problems

The first task in the study of the multidimensional conservation laws is to establish the local theory of classical solutions. The main tool in this stage is

the energy estimates. In most cases the existence and the stability of solutions to partial differential equations rely on various energy estimates. The main estimates for multidimensional system of conservation laws are measured by Sobolev norms. The usual Sobolev space is good in treating Cauchy problems and boundary value problems when the boundary is not characteristic. However, when the boundary is characteristic, the derivatives of the system on the normal direction at the boundary are not "complete". Hence it is inevitable to meet the smoothness loss at the boundary when a linearization or an iterative process is designed to solve a boundary value problem for nonlinear equations with characteristic boundary. On the other hand, the characteristic boundary appears very often in the systems of conservation laws. For instance, if one consider an inviscid flow in a domain with a rigid wall as its boundary (or a part of its boundary), then the impermeable condition let the boundary be characteristic for the Euler system describing the inviscid flow.

People found that in the proof of the regularity of the possible solutions (or approximate solutions) to the characteristic boundary value problems for multidimensional systems of conservation laws, it is necessary to introduce a new weighted Sobolev space to deal with these problems (e.g. see [9]). This space needs to have such a property: "one order gain of differentiation in normal direction to the boundary should be compensated by two orders loss of differentiation in tangential directions". More precisely, let $D_q^p u$ denote the derivative of the function u, where p is the order in tangential directions, q is the order in the normal direction, let H_t^s denote the Sobolev space of u, satisfying $D_q^p u \in L^2([0,h]\times\Omega)$ $(p \le s, q \le t)$. Define

$$B_p = \bigcap_{d \le \frac{p}{2}} H_d^{p-2d}([0,h]\times\Omega), \tag{3}$$

which is the Sobolev space describing the above-mentioned property (the space is also denoted by H_*^p in [38],[41] etc.). Such a space has been employed to prove the existence and uniqueness of the local existence of solutions to characteristic boundary value problems [9],[42]. The Sobolev spaces with such a property is also applied to other physical problems involving characteristic boundaries (for instance, [2],[10]).

On characteristic boundary value problems of symmetric hyperbolic systems we would like also mention the contributions given by R.Agemi[1], D.G.Ebin [25] and P.Secchi [41].

3. 1-D Like Problems with M-D Perturbation – Fan-shaped Wave Structure

Like the nonlinear systems of conservation laws in one space-dimension a general solution of a nonlinear multidimensional system of conservation laws may

develop singularities, no matter how smooth the initial data are (e.g. see [43]). Therefore, in the next stage people must study the theory of weak solutions. The main nonlinear waves for multidimensional systems of conservation laws are shock wave, simple wave and contact discontinuity, like in the one-space-dimensional case. However, the structure of nonlinear waves is much more complicated. It should be noticed that the front of nonlinear waves often plays the role of the boundary of a domain where the solution is smooth. Generally, the location of the wave front should be determined together with the solution. Hence to find the weak solution involving various nonlinear waves often require to solve a free boundary value problem.

Due to the complexity of characteristic varieties of multidimensional systems the nonlinear wave structure for these systems is plentiful. The simplest case is the 1-d like structure with m-d perturbations. In 1-d case, besides a single wave propagates along a curve, one often meets such a wave structure: several waves represented by a set of curves issue from a point. Similarly, a common wave structure in multidimensional case is that several nonlinear waves issue from a smooth curve. Such a wave structure is called **fan-shaped wave structure**. For instance, consider a Cauchy problem of multidimensional system of conservation laws with initial data, which is discontinuous along an smooth curve (for simplicity we only consider two-space-dimensional case here). Then there will possibly be shock waves, simple waves and contact discontinuities issuing from the curve bearing the discontinuity of the initial data. Especially, under some restrictions on the initial data the solution will only contain one wave among these three kinds of nonlinear waves.

In 1983 A.Majda started the study of weak solutions to multidimensional system of conservation laws with fan-shaped wave structure. He first proved the existence of the solution to the initial value problem of a nonlinear multidimensional system involving a shock front, which issues from a curve carrying the discontinuity of the initial data ([33]). In his work the Cauchy problem with discontinuous data is treated as a free boundary value problem, and the shock front is the free boundary, which is to be determined together with the solution. In the free boundary value problem the Rankine-Hugoniot conditions give a differential relation of the function describing the unknown shock front. An important fact is that the relation is an elliptic differential system for the unknown function on the free boundary. Meanwhile, as a boundary value problem for the original system of conservation laws the uniform Kreiss-Lopatinski condition on the boundary is fulfilled. Such a property let suitable estimates, which dominates the variation of the solution with a shock front, can be established. Then the estimates directly lead to the stability of the solution to the linearized problem and the existence of the local solution of the nonlinear problem near the curve describing the discontinuity of the initial data.

Under some other restrictions on the initial data one can find a solution to the Cauchy problem containing a centered rarefaction wave. In this case the

rarefaction wave is formed by a family of characteristic surfaces in two-space-dimensional case. All these characteristic surfaces issue from a curve carrying the discontinuity of the initial data. The rarefaction wave likes a fan with a front surface and a back surface. The solution is differentiable inside the rarefaction wave region, and it is only continuous on the front surface and the back surface of the fan. Moreover, the solution is discontinuous at the edge of the fan. The front surface and the back surface are unknown and has to be determined with the solution together, hence they can also be regarded as free boundaries. Different from the shock front solution case the front and the back of the rarefaction wave are characteristics. Hence in the corresponding iterative process to establish the existence of the solution to the nonlinear problems one will also confront the "derivative loss" difficulty, which happens even for fixed boundary value problems with characteristic boundary, as indicated in the previous section. An delicate treatment is given in [2], where the Nash-Moser iterative scheme is applied to overcome this difficulty. Meanwhile, the weighted Sobolev space B_p defined in the previous section is also employed once more. In [2] the local existence of the solution with a rarefaction wave are proved.

The most difficult case in solving Cauchy problem of multi-dimensional system of conservation laws with discontinuous initial data is the case of contact discontinuity, which is a surface formed by stream lines, and is also called compressible vortex sheets. This surface is obviously characteristic. However, different from the rarefaction wave case the solution on this characteristic surface is also discontinuous. On the other hand, different from the shock front case the uniform Kreiss-Lopatinskii condition on the boundary is not satisfied. In other words, in the study of the compressible vortex sheets one will confront the difficulties, which appear in either the shock front case or the rarefaction wave case. J-F.Coulombel and P.Secchi studied the vortex sheets case for multidimensional system of conservation laws in [23]. They noticed that the normal component of the unknown vector on the boundary satisfies the weak Lopatinskii conditions. When the non-degenerate part of the Lopatinskii condition is extracted, the microlocal symmetrizer can then be constructed. In the meantime, the Nash-Moser iterative scheme is also applied to avoid the derivatives loss. These techniques help them to derive the energy estimates, which ensure the stability of solutions to linearized problems and the existence of solutions to nonlinear problems.

When the initial data do not satisfy the restrictions given in the above three special cases the solution to the multidimensional systems under consideration may contain more nonlinear waves, like two shocks (see [35],[37]), one shock and one rarefaction wave (see [29]) etc. General data with discontinuity on a smooth curve may develop all three kinds of nonlinear waves (shock, rarefaction wave and contact discontinuity). Although the main difficulties in the three individual cases have been overcome, the result in most general cases has not been established so far. Obviously, such a result is significant and is anticipated.

People may also find fan-shaped configurations of nonlinear waves in many other problems in fluid dynamics. For instance, an important problem in gas dynamics is the study of supersonic flow past a wedge. For a steady supersonic flow past a three-dimensional wedge, when the attack angle and the vertex angle of the wedge are well controlled, an attached shock front at the edge of the wedge will be formed. Such a physical problem can also be reduced to a boundary value problem in a domain between the attached shock front and the surface of the wedge. The shock front is the free boundary for this boundary value problem. The local existence of the solution with the attached shock front was proved in [11]. Other results on physical problems with fan-shaped wave configurations can be found in [10] for shock reflection by a smooth surface, in [36] for propagation of sound waves etc.

4. Essentially M-D Problems – Flower-shaped Wave Structure

It should be emphasized that in the multidimensional space many complicated wave structure are not 1-d like ones with m-d perturbations. Such structures are essentially multidimensional. Based on the progress in the study of various multidimensional problems with fan-shaped wave structure the researchers gradually concentrate their concerns on more complicated cases.

A good example of the essentially multidimensional problem is the problem on supersonic flow past a pointed body. Like the problem on supersonic flow past a wedge, when the vertex angle of the pointed body is less than a critical value, the shock front is attached at the tip of the body, forming a bigger conical surface. Obviously, such a shock front is not a perturbation of a plane shock, and the state between the shock and the surface of the body is not a perturbation of a constant state either. Here the shock front, as well as the surface of the body, issues from a single point – the tip of the conical body. Hence such a wave structure is called **flower-shaped wave structure**. We notice that the domain is formed by two conical surfaces with strong singularity at the tip, which will cause new difficulties.

In [12] the author gives a proof of the existence of shock front solution near the tip. The problem is first approximated by the straight version of the original problem, i.e. the pointed body is replaced by a conical body with straight generating lines, and the coming flow is assumed to be constant. For the problem of straight version one can make analysis in self-similar coordinates, by which the problem can be reduced to a free boundary value problem of an elliptic equation.

The main result obtained in [12] is

Theorem 4.1. *Assume that a pointed body is given by* $r = b(z, \theta)$*, where* (r, θ, z) *is the relative cylindrical coordinates,* $r = R/z$ *is the ratio of the regular*

cylindrical coordinates R and z, $b(z,\theta)$ is a small perturbation of a constant b_0 in the sense of [12]. Assume that a supersonic flow parallel to the z-axis comes from infinity with speed $q = q_\infty$ satisfying $q_\infty > a_\infty \left(= \left(\frac{\gamma p_\infty}{\rho_\infty}\right)^{\frac{1}{2}}\right)$, where p_∞, ρ_∞ are the pressure and the density at infinity respectively. Besides, b_0 is less than a critical value determined by $q_\infty, p_\infty, \rho_\infty$ introduced in [24]. Then the problem of the supersonic flow past the pointed body admits a local weak entropy solution with a pointed shock front attached at the origin.

The result on the existence of the solution with its shock front structure near the tip of the body enable us to study global existence and the asymptotic behavior of the flow behind the shock waves [22],[31],[47].

Another well-known problem involving flower-shaped wave structure is "shock reflection by a wedge". This is a problem on unsteady flow in two-dimensional space.

When a plane shock front hits a wedge, then a reflected shock will move outward from the edge of the wedge, while the incident shock moves forward in time. By symmetry the problem amounts to consider a shock front hitting a ramp (hence the problem is also called "shock reflection by a ramp"). Let the instant, when the shock front touches the edge of the wedge, be the time $t = 0$, the problem is invariant under the dilation of the time coordinate and the space coordinates. Hence one can look for the self-similar solution of the problem. The corresponding flow in the self-similar coordinates $\xi = x/t, \eta = y/t$ is called **pseudo-steady flow**, because all parameters of the flow depend only on the coordinates (ξ, η), and does not depend on the time t explicitly.

Since all possible waves in this problem are invariant under the dilation, they can be viewed as a "flower" generated from the origin. Hence in the (t, x, y) physical space we obtain a flower-shaped wave structure. Depending on the vertex angle of the wedge (or the angle between the ramp and the horizon) the shock reflection may have various patterns. Among them the simplest case is the **regular reflection**, for which only a smooth curved reflected shock moves outward like an expanding bubble. Since in the region behind the reflected shock both relatively supersonic flow and relatively subsonic flow will occur, to prove the existence of the solution to the regular shock reflection problem needs to solve a nonlinear mixed type equation, or at least a nonlinear degenerate elliptic equation. B.L.Keyfitz and her collaborators [5],[6] first use UTSD (unsteady transonic small disturbance) equation as the model to establish a result on the existence of solution to the shock reflection problem. Later, G.Q. Chen and M.Feldman [8] use the potential flow equation as the model to discuss regular reflection. Since the coefficients of the potential flow equation depend on the derivatives of the unknown function (the gradient of the flow potential), the proof for the latter case is more difficult. Indeed, the authors of [8] indicated the following conclusion:

For a given supersonic incoming flow, one can find a suitable angle θ_c and a number α, such that, if the angle of the inclination θ_w of the ramp is in $\left(\theta_c, \frac{\pi}{2}\right)$, then there is a global self-similar solution of the potential flow equation, satisfying the assigned boundary conditions. The solution is globally in $C^{1,\alpha}$, and is C^∞ outside of the shock front and the sonic line. Moreover, the solution is stable with respect to the change of the angle and converges to the normal reflection as $\theta_w \to \pi/2$.

For some combinations of parameters of the upstream flow and the angle of the ramp the regular reflection is impossible. The corresponding wave patterns in these cases are called irregular shock reflection. Among them the most important case is the **Mach reflection**, which is composed of three shock fronts (incident, reflected and Mach stem) and a contact discontinuity. Near the intersection of these waves the above wave configuration is called **Mach configuration**. It is interesting that in the self-similar coordinate plane we again confront such a wave structure, in which several nonlinear waves issue from a point.

Even for the steady compressible flow the shock reflection can also be distinguished as regular reflection (oblique shock reflection) and irregular reflection (including Mach reflection). In fact, it is von Neumann, who first found the wave structure in Mach reflection and proposed the concept of Mach configuration in 1943 based on the numerous physical experiments and mathematical analysis for the shock reflection in steady compressible flow [3], [39].

In the study of Mach configuration, if all shock fronts and the contact discontinuity are straight lines and the states in each domains separated by these nonlinear waves are constant, the configuration is called **flat Mach configuration**. Locally at the triple point, the Mach configuration always can be approximately viewed as flat configuration. The related problem, which people are deeply concerned with, is the stability of Mach configuration, because only stable wave configuration can actually occur in physics.

Generally, the flow behind the Mach stem is always subsonic, but the flow passing across the incident shock and then the reflected shock can be either subsonic or supersonic. Therefore, referring to the flow in the downstream part we classify the Mach configurations as E-E type and E-H type. For E-E type Mach configuration the flow in the downstream part is composed of two branches of subsonic flow separated by a stream line. For E-H type Mach configuration the flow in the downstream part is composed of a supersonic flow and a subsonic flow adjacent to each other with a stream line separating them.

In [14], [15] we proved the stability of the E-E type Mach configuration for both steady case and unsteady case. For example, the result in the steady case is:

Theorem 4.2. *Assume that the constant states $U_i^0 (0 \le i \le 3)$, the shock fronts $S_i (i = 1, 2, 3)$ and the contact discontinuity D form a flat Mach configuration, where U_0^0 is the state of the coming supersonic flow, U_1^0 is the state behind*

the incident shock S_1, U_2^0, U_3^0 are the states behind the Mach stem S_2 and the reflected shock S_3 respectively. U_2^0 and U_3^0 are both subsonic and are separated by a contact discontinuity D. Assume that U_0 and $\widetilde{S_1}$ are non-flat perturbations of U_0^0 and S_1, then one can find a non-flat Mach configuration in a neighborhood of the triple point, where the shock fronts S_2, S_3 and the contact discontinuity D are slightly perturbed. Correspondingly, the states U_1, U_2, U_3 in each domain separated by these nonlinear waves are the perturbation of the corresponding states for the original flat Mach configuration.

The stability of the E-H type Mach configuration is also proved in [20] under an additional assumption that the reflected shock is weak.

The above result supported the reasonableness of the Mach configuration proposed by von Neumann [39]. However, the global existence of the Mach reflection is still quite open.

V.Elling and T.P.Liu studied the problem on a moving wedge hitting a static gas in [26]. Suppose there is a uniform static gas filling up the whole space outside a given wedge. The wedge suddenly moves into the air with a constant speed in the direction of its symmetric axis, then the gas flow caused by the motion of the wedge is also self-similar. Obviously, in the time-space coordinate system the wave configuration is flower-shaped. If the speed of the wedge is supersonic, then near the head of the wedge the motion of the air is not influenced by the initial state, so that there is a shock front attached at the edge of the wedge, provided the vertex angle of the wedge is less than a critical value. Meanwhile, far from the vertex of the wedge, the motion of the air amounts to one dimensional: a plane wall moves into the domain filled with static gas along the normal direction. According to the theory of one-dimensional conservation laws there will be a plane shock, moving along the direction normal to he plane wall. On the self-similar coordinate plane the latter is called tail shock. The straight head shock attached at the edge and the straight tail shock are connected by a curved shock. In accordance, behind the shock the flow is a mixture of a relatively supersonic flow and a relatively subsonic flow. The existence of the solution for this problem was established in [26].

Another interesting physical problem involving flower-shaped wave structure is the dam-collapse problem [44]. Assume that a wedge-shaped reservoir is filled with water. At the time $t = 0$ the dam suddenly collapsed so that the water floods outside of the reservoir. The problem is to determine the flow in $t > 0$. Since the motion of the water is governed by the shallow water equation, which is quite similar to the Euler equation in gas dynamics, the dam-collapse problem is also similar to the problem of expansion of gas contained in a wedge-shaped domain into vacuum. In both cases the motion of the fluid is given by an interaction of two rarefaction waves indeed. The problem was solved in [32].

More complicated wave structures with flower-shaped wave configuration appear in the study of multidimensional Riemann problems. Consider

two-dimensional Riemann problem, which is a Cauchy problem of the two-dimensional system of conservation laws with piecewise constant initial data, which takes different constants in different angular domains. In [49] the various cases for 2-d Riemann problems are mentioned and classified, where the initial data take different constants in four quadrants. It is well known that the Riemann problems in one-space-dimensional case has been well studied. Particularly, the results on 1-d Riemann problem play the fundamental role in the theory of conservation laws (see [27]). However, the study on Riemann problems in multidimensional case is only at its beginning.

The Riemann problem is invariant under the dilation of the coordinates. Its solution also has flower-shaped wave configuration. By using self-similar coordinates all waves becomes fixed in the new coordinate plane. Hence a d+1 dimensional unsteady problem (1 time-dimension plus d space-dimensions) becomes a d-dimensional problem in self-similar coordinate system. The latter is usually called pseudo-steady problem. Therefore, to determine a flower-shaped wave structure for d+1 dimensional unsteady problem is then reduced to look for a global solution on the self-similar coordinate system. Far away from the origin the influence of the origin vanishes, so that the d+1 dimensional unsteady problem is one space-dimensional, which can be solved by using the theory of one-dimensional system of conservation laws. In accordance, for a given multi-dimensional Riemann problem one may have many nonlinear waves (formed by straight characteristics) coming from infinity in different directions, and these nonlinear waves will interact when they meet together. The plentiful phenomena of interaction of these waves lead to the great complexity of the nonlinear wave structure either in the self-similar coordinates or in the original physical coordinates.

The Riemann problem is a special initial value problem for the hyperbolic system of conservation laws. Like the setting of the Riemann problems we can also consider some initial boundary value problems invariant under dilation of time coordinate and space coordinates. Many physical problems, including the above-mentioned "shock reflection by a ramp" and "dam-collapse", can be derived in such a way, that the initial data take different constants in different sectors, while some sectors are solid, where no flow could go into. Such problems are called **initial boundary value problems of Riemann type**. For instance, we can take initial data as follows. The whole plane is separated by the rays $\theta = \theta_0$ $(0 < \theta_0 < \pi/2)$, $\theta = \pi/2$ and $\theta = -\pi$ to three sectors. The sector $-\pi < \theta < \theta_0$ are solid and no gas can go into. Meanwhile, the gas is assumed to take different constant states in $\theta_0 < \theta < \pi/2$ and $\pi/2 < \theta < \pi$. Moreover, the flow parameters in both sides of $\theta = \pi/2$ can determine a single plane shock moving forward to the ramp $\theta = \theta_0$. Then, the initial boundary value problem with such data amounts to the physical problem "shock reflection by a ramp".

Like the above setting one can also discuss other initial boundary value problems of Riemann type. For instance, in the above example, if the flow

parameters on the both side of $\theta = \pi/2$ can determine a single rarefaction wave moving forward to the ramp, then we obtain a problem "reflection of rarefaction wave by a ramp". Similarly, if $\theta_0 < 0$ we can obtain the problem "shock diffraction by a convex angle". In the latter case there may not be any reflected shock, but there is a sonic wave propagating from the origin to infinity.

The dam-collapse problem can also be considered as such an initial boundary value problem of Riemann type. The problem corresponds to the case with initial data: the gas takes non-vacuum constant state in a sector $-\theta_0 < \theta < \theta_0$, while the domain $\theta_0 < \theta < 2\pi - \theta_0$ at the initial time is vacuum. Similarly, one can also consider the case corresponding to the initial data: the gas takes non-vacuum state in the domain $\theta_0 < \theta < \pi/2$, while the domain $\pi/2 < \theta < \pi$ is vacuum and the domain $\pi < \theta < 2\pi + \theta_0$ is solid. The problem will give a wave pattern of the reflection of a "full" rarefaction wave (expanding up to vacuum) by a ramp.

5. Global Theory and Mixed Type Equations

The previous section shows that in the study of compressible flow many problems involve both supersonic flow and subsonic flow (or relatively supersonic flow and relatively subsonic flow in pseudo-steady case), then the system describing the flow has complex characteristics in subsonic region, while its all characteristics in supersonic region are real. Therefore, in order to study the flow globally, we have to consider transonic flows and mixed type equations.

The study of mixed type equations was initiated by F.Tricomi in 1923. The Tricomi equation $yu_{xx} + u_{yy} = 0$ was named by his successors for his contributions to this area [45]. Later, the mixed type equations $u_{xx} + yu_{yy} = 0$ and $u_{xx} + \mathrm{sgn}y\ u_{yy} = 0$, called Keldysh equation and Lavrentiev-Bitsadze equation respectively, are also proposed and studied by many authors. Both the Tricomi equation and the Keldysh equation are degenerate on the line where the type of the equation is changed. The difference is that for the Tricomi equation the characteristics in hyperbolic region is perpendicular to the degenerate line, while for the Keldysh equation the characteristics in hyperbolic region is tangential to the degenerate line. On the other hand, the Lavrentiev-Bitsadze equation has discontinuous coefficients with discontinuity on the line, where the equation changes its type. These three equations are prototypes of more complicated mixed type equations arisen in various physical problems, particularly arisen in fluid dynamics. Therefore, the study on them are important for the development of the theory of fluid dynamics. However, the change of type often causes great difficulties in the corresponding study.

Next we introduce some problems in gas dynamics related to the mixed type equations, most of them are still open. Obviously, the solution of them will promote a series of new progress in both mathematics and physics. We believe that the progress on the study of mixed type equations will bring us a breakthrough in the study of multidimensional system of conservation laws.

The first example is the E-H type Mach configuration. As mentioned in the previous section, in the discussion of Mach reflection the flow behind the reflected shock could be supersonic, which is adjacent to a subsonic flow with a slip line separating them. Such a wave structure is called E-H type Mach configuration. To determine such a (non-flat) wave configuration near the triple point one has to solve a free boundary value problem of nonlinear mixed type equation. On the curve, where the equation changes its type, the solution is continuous, while its derivatives should satisfy some consistency conditions. Since the coefficients of the equation is discontinuous due to the discontinuity of the flow parameters on the contact discontinuity, the equation belongs to Lavrentiev-Bitsadze's mixed type equation. On the other hand, the data on boundary conditions are assigned on the whole boundary of the elliptic region and on a part of the boundary of the hyperbolic region. Such a setting of boundary conditions is similar to that for Tricomi problem [45]. In [20] the local existence and stability for E-H Mach configuration is proved under the assumption that the reflected shock is weak.

Many hyperbolic problems in physical time-space variables may lead to problems for mixed type equations in self-similar coordinates. Indeed, in the self-similar coordinates plane the equation (2) can be written as

$$(c^2 - (\phi_\xi - \xi)^2)\phi_{\xi\xi} - 2(\phi_\xi - \xi)(\phi_\eta - \eta)\phi_{\xi\eta} + (c^2 - (\phi_\eta - \eta)^2)\phi_{\eta\eta} = 0, \quad (4)$$

where ϕ is the potential of the velocity of the flow, $(\phi_\xi,\ \phi_\eta) = (u, v)$ is the velocity of the flow. The discriminant of the equation is $c^2(c^2-(u-\xi)^2-(v-\eta)^2)$, so that the equation is elliptic for (ξ, η) satisfying $c^2 < (u-\xi)^2+(v-\eta)^2$, and is hyperbolic for (ξ, η) satisfying $c^2 > (u-\xi)^2 + (v-\eta)^2$. Generally, the equation is elliptic as $(\xi, \eta) = (u, v)$, while it is hyperbolic as $(\xi, \eta) \to \infty$. Therefore, the equation is of mixed type generally.

The change of type causes much difficulties in various problems related to mixed type equations. While in some exceptional cases we can consider the solution in the hyperbolic region or in the elliptic region separately. One exceptional case is that the flow in the hyperbolic part is constant, and can be determined by only solving some algebraic equations. Then the problem for the mixed type equation will be reduced to a boundary value problem for a degenerate elliptic equation. The problem on the regular shock reflection or the interaction of two shock fronts belongs to such cases. By taking Chaplygin gas as model to describe the compressible flow, D.Serre studied the interaction of multidimensional shocks. The problem is finally reduced to a Dirichlet boundary value problem for a degenerate elliptic equation of second order, and the unique existence of its global solution is proved (see [40]). Meanwhile, the problems studied in [8], [26] are also reduced to a boundary value problem for degenerate elliptic equations of second order. The dam-collapse problem is another exceptional case. In this case the elliptic domain in the problem is reduced to a single point, so that the mixed type equation is reduced to a degenerate hyperbolic equation. However, in general case the interaction of

the solution in the hyperbolic region and the solution in the elliptic region is inevitable.

Next we give two interesting problems related to mixed type equations in steady compressible flow. The first one is the flow in de Laval nozzle. A de Laval nozzle contains a converging part near the entrance and a diverging part near the exit. The nozzle has a throat in the middle, where the cross sectional area takes minimum. It is known that a compressible subsonic flow will speed up as the nozzle becomes narrow and it will slow down as the nozzle becomes wide. In contrary, a compressible supersonic flow will slow down in a convergent part of a nozzle and will speed up in a divergent part of a nozzle. Therefore, the subsonic flow at the entrance of the de Laval nozzle will speed up as the nozzle is getting narrow. As Courant-Friedrichs conjectured in [24], for a suitable incoming subsonic flow and an assigned pressure at the exit the whole flow pattern could be as follows: the flow reaches sonic speed near the throat, then after passing over the throat the flow becomes supersonic and is accelerating further. Afterwards, if the pressure or other flow parameters at the exit are well controlled the supersonic flow may passes across a transonic shock front and becomes a subsonic flow again, which finally reaches the assigned pressure at the exit. Then the problem is how to determine the flow in the whole nozzle, as well as the location of the sonic line and the possible transonic shock, provided one only knows the incoming flow at the entrance and the assigned condition at the exit.

Many works based on multidimensional PDE analysis for such a problem are proceeded (see [7], [16], [17], [30], [46]). A recent result is that for an two-dimensional expanding nozzle if the supersonic flow at the entrance is given and the pressure at the exit is suitably controlled, then the transonic shock front and the subsonic flow between the shock front and the exit can be uniquely determined [17]. The result coincide with the Courant-Friedrichs' conjecture [24] in the divergent part of the de Laval nozzle. However, how does a subsonic flow continuously transforms to a supersonic flow under the influence of the shape of the nozzle is still open. Obviously, the complete result on the existence and stability of the flow in de Laval nozzle depends on the study of mixed type equations. We notice that at least some characteristics of the equation in the hyperbolic region is tangential to the sonic line, then the nonlinear mixed type equation describing the compressible flow in the de Laval nozzle may have more similarity to the Keldysh equation.

The second example is the supersonic flow past a blunt body. It is also a long standing problem in gas dynamics (see [24], [18]). When a uniform supersonic flow attacks a fixed blunt body, there will appear a detached shock front ahead of the body. Near the head of the body the shock is almost normal, so that the flow behind the shock must be subsonic. On the other hand, if the blunt body is finite, then the flow passes around the blunt body will finally merge together. Therefore, far away from the head of the body the angle between the shock front and the voticity of the flow will gradually become small, so

that eventually the flow behind the shock front will be supersonic. The above analysis indicates that the steady flow in the whole region between the shock front and the surface of the body must be transonic. The problem to determine the flow and the location of the shock front is also a free boundary value problem for a nonlinear mixed type equation. In the meantime, the global existence of solution must be considered. So far the analytical study on this problem is still formidable and complete open.

In the end of this paper we would like especially emphasize the importance of the study of mixed type equations. The theory of mixed type equations is much less mature than the theory of elliptic equations and hyperbolic equations. The new theory and technique to deal with various boundary value problems of mixed type equations (particularly, nonlinear mixed type equations) are crucial to the development of multidimensional conservation laws. They will also bring a breakthrough to the whole theory of partial differential equations.

We certainly have not mentioned all difficulties in the study of multidimensional systems of conservation laws. Particularly, the influence of vorticity has not been discussed. Evidently, it is a troublesome factor in the study of various physical problems. Besides, the study on viscous multidimensional conservation laws has also not been discussed. People probably have to combine the study of viscous conservation laws and Navier-Stokers equations to get better understanding on the role and influence of viscosity and vorticity.

References

[1] R.Agemi, *The initial boundary value problem for inviscid barotropic fluid motion*, Hokkaido Math. J., 10(1981), 156–182.

[2] S.Alinhac, *Existence d'ondes de rarefaction pour des systèmes quasi-linéaires hyperboliques multidimensionnels*, Comm. P.D.E., 14(1989), 173–230.

[3] G.Ben-Dor, *Shock Waves Reflection Phenomena (second edition)* Springer-Verlag, Berlin, Heiderberg, New York, 2007.

[4] L.Bers, *Mathematical Aspects of subsonic and transonic gas dynamics*, John Wiley & Sons, New York, 1958.

[5] S.Canic, B.Keyfitz and G.M Lieberman, *A proof of existence of perturbed steady transonic shocks via a free boundary problem*, Comm. Pure Appl. Math., 53(2000), 484–511.

[6] S.Canic, B.Keyfitz and E.H.Kim, *A free boundary problem for a quasilinear degenerate elliptic equation: regular reflection of weak shocks*, Comm. Pure Appl. Math., 55(2002), 71–92.

[7] G.Q.Chen and M.Feldman, *Multidimensional transonic shocks and free boundary problems for nonlinear equations of mixed type*, Jour. Amer. Math. Soc., 16(2003), 464–491.

[8] G.Q.Chen and M.Feldman, *Potential theory for shock reflection by a large-angle wedge*. Proceedings of the National Academy of Sciences of USA., 102(2005),

15368–15372. Global solution to shock reflection by large-angle wedges for potential flow. Ann. Math., accepted 2006.

[9] S.X.Chen, *On the initial-boundary value problems for the quasilinear symmetric hyperbolic system with characteristic boundary*, Chin. Ann. Math., 3(A)(1982), 223–232 (Chinese version), Front, Math. China, 2(2007), 87–102 (translated English version).

[10] S.X.Chen, *On reflection of multidimensional shock front*, Jour. Diff. Eqs., 80(1989), 199–236.

[11] S.X.Chen, *Existence of local solution to supersonic flow around a three dimensional wing*, Adv. Appl. Math., 13(1992), 273–304.

[12] S.X.Chen, *Existence of stationary supersonic flows past a pointed body*, Arch. Rat. Mech. Anal., 156(2001), 141–181.

[13] S.X.Chen, *A singular multi-dimensional piston problem in compressible flow*, Jour. Diff. Eqs., 189(2003), 292–317.

[14] S.X.Chen, *Stability of a Mach Configuration*, Comm. Pure Appl. Math., 59(2006), 1–33.

[15] S.X.Chen, *Mach configuration in pseudo-stationary compressible flow*, Jour. Amer. Math. Soc., 21(2008), 63–100.

[16] S.X.Chen, *Transonic in 3-D compressible flow passing a duct with a general section for Euler systems*, Trans. Amer. Math. Soc., 360(2008), 5265–5289.

[17] S.X.Chen, *Compressible flow and transonic shock in a diverging nozzle*, Comm. Math. Phys., 289(2009), 75–106.

[18] S.X.Chen, *Study on Mach reflection and Mach configuration*, Proceedings in Applied Mathematics, Hyperbolic Problems: Theory, Numerics and Applications, American Mathematical Society, 67(2009), 53–71.

[19] S.X.Chen, *Mixed type equation in Gas Dynamics*, to be published in Quat. Appl. Math.

[20] S.X.Chen, *E-H type Mach configuration and its stability*, (to appear).

[21] S.X.Chen and Dening Li, *Supersonic flow past a symmetric curved cone*, Indiana Univ. Math. Jour., 49(2000), 1411–1435.

[22] S.X.Chen, Z.P.Xin and H.C.Yin, *Global shock waves for the supersonic flow past a perturbed cone*, Comm. Math. Phys., 228(2002), 47–84.

[23] J-F.Coulombel and P.Secchi, *Nonlinear compressible vortex sheets in two space dimensions*, Ann. Sci. Ec. Norm. Super., 41(2008), 85–139.

[24] R.Courant and K.O.Friedrichs, *Supersonic Flow and Shock Waves*, Springer-Verlag, New York, 1949.

[25] D.G.Ebin, *The initial-boundary value problem for subsonic fluid motion*, Comm. Pure Appl. Math., 32(1979), 1–19.

[26] V. Elling and T. P. Liu, *Supersonic flow onto a solid wedge*, Comm. Pure Appl. Math., 61(2008), 1347–1448.

[27] P.D.Lax, *Hyperbolic system of conservation laws*, Comm. Pure Appl. Math., 10(1957), 537–566.

[28] P.D.Lax, *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, Conf. Board Math. Sci., 11, SIAM,(1973).

[29] D.Li, *Rarefaction and shock waves for multi-dimensional hyperbolic conservation laws*, Comm. in PDEs, 16(1991), 425–450.

[30] Jun Li, Zhouping Xin and Huicheng Yin, *On transonic shocks in a conic divergent nozzle with axi-symmetric exit pressures*, Jour. Diff. Eqns., 248(2010), 423–469.

[31] W.C.Lien and T.P.Liu, *Nonlinear stability of a self-similar 3-D gas flow*, Comm. Math. Phys., 304(1999), 524–549.

[32] J.Q.Li and Y.Zheng, *Interaction of rarefaction waves of the two-dimensional self-similar Euler equations*, Arch. Rat. Mech. Anal., 193(2009), 623–657.

[33] A.J.Majda *The stability of multi-dimensional shock front, The existence of multi-dimensional shock front*, Memoirs Amer. Math. Soc., 275–281(1983).

[34] A.J.Majda, *One perspective on open problems in multi-dimensional conservation laws*, IMA Math. Appl. 29(1991), 217–237.

[35] G.Metivier, *Interaction de deux chocs pour un systeme de deux lois de conservation en dimension deux d'espace*, Trans. Amer. Math. Soc., 296(1986), 431–479.

[36] G.Metivier, *Ondes Soniques*. Jour. Math. Pure Appl., 70(1991), 197–268.

[37] M.Mnif, *Probleme de Riemann pour une loi conservation scalaire hyperbolique d'order deux*, Comm. in PDEs, 22(1997), 1589–1627.

[38] A.Morando, P.Secchi and P.Trebeschi, *Regularity of solutions to characteristic initial-boundary value problems for symmetrizable systems*, Jour. Hyp. Diff. Eqns, 6(2009), 753–808.

[39] J. von Neumann, *Oblique reflection of shocks*, PB-37079. U.S.Department of Commerce, Office of Technical Services, Washington D.C., 1943.

[40] D.Serre, *Multidimensional shock interaction for a Chaplygin gas*, Arch. Rat. Mech. Anal., 191(2009), 539–577.

[41] P.Secchi, *Well-posedness of characteristic symmetric hyperbolic systems*, Arch. Rat. Mech. Anal., 134(1996), 155–197.

[42] M.Ohno, Y.Shizuta and T.Yanagisawa, *The initial boundary value problem for linear symmetric hyperbolic system with characteristics of constant multiplicity*, Jour. Math. Kyoto Univ., 35(1995), 143–210.

[43] T.Sideris, *Formation of singularities in three-dimensional compressible fluids*, Comm. Math. Phys., 101(1985), 475–485.

[44] V.A.Suchkov, *Flow into a vacuum along an oblique wall*, Jour. Appl. Math. Mech., 27(1963), 1132–1134.

[45] F.Tricomi, *Sulle equazioni lineari alle derivate parziali di secondo ordino, di tipo misto*, Rendi. Reale Accad. Lincei, 14(1923), 134–247.

[46] Z.P.Xin and H.C.Yin, *Transonic shock in a nozzle I:two–dimensional case*, Comm. Pure Appl. Math., 58(2005), 999–1050.

[47] Z.P.Xin and H.C.Yin, *Global multidimensional shock wave for the steady supersonic flow past a three-dimensional curved cone*, Anal. Appl.(Singap.)4(2006), 101–132.

[48] Y.Zheng, *System of Conservation Laws. Two-dimensional Riemann Problems*, Birkhäuser, Boston, 2001.

[49] T.Zhang and Y.Zheng, *Conjecture on the structure of solution of the Riemann problem for two-dimensional gas dynamics systems*, SIAM Jour. Math. Anal., 21(1990), 593–619.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Finite Morse Index and Linearized Stable Solutions on Bounded and Unbounded Domains

E. N. Dancer*

Abstract

We discuss stable and finite Morse index solutions of nonlinear partial differential equations. We discuss problems on all of space, on half spaces and on bounded domains where either the diffusion is small or the solutions are large.

Mathematics Subject Classification (2010). Primary 35J61; Secondary 35J91.

Keywords. Nonlinear elliptic equations, stable solutions, finite Morse index solutions.

In this lecture, I wish to discuss solutions which are stable or not too unstable of problems

$$\begin{aligned} -\epsilon^2 \Delta u &= f(u) \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{1}$$

where Ω is a smooth bounded domain in $\mathbb{R}^N$ or $\Omega = \mathbb{R}^N$ or Ω is a half space. We could also consider other boundary conditions. We usually take ϵ small or $\epsilon = 1$. In fact we are usually interested in the case of a bounded domain. The half space or whole space problems occur as limit problems for this problem. We always assume f is C^1. To see why this should be the case, we assume $0 \in \Omega$ (which we can do by a translation) and then we change the variables $X = \epsilon^{-1}x$ (which is a large stretching of the variables). In this case in our new variables, our equation becomes

$$\begin{aligned} -\Delta' u &= f(u) \quad \text{in } \epsilon^{-1}\Omega \\ u &= 0 \quad \text{on } \partial(\epsilon^{-1}\Omega) \end{aligned} \tag{2}$$

*School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia. E-mail: norman.dancer@sydney.edu.au.

where Δ' is the Laplacian in the new variables. Since $0 \in$int Ω, then, if ϵ is small, $\epsilon^{-1}\Omega$ contains larger and larger balls and hence it is natural to expect that as $\epsilon \to 0$ the solutions of (1) approach the solutions of (2) on $\mathbb{R}^N$. This is in fact quite easy to prove if u is uniformly bounded. (Moreover many properties of the solutions persist in the limit.) Note that it may happen that $|u|$ has its maximum within order ϵ of the boundary. In this case we choose our origin of coordinates to be the point of $\partial\Omega$ closest to the maximum point $|u|$ and stretch the coordinates much as before. The stretching of coordinates tends to flatten the boundary (where we used $\partial\Omega$ is smooth) and it turns out that the limit problem is the half space problem (after a rotation of axis). Thus we see that the half spaces and $\mathbb{R}^N$ problems occur naturally when we consider bounded domain problems. We refer to using these limiting arguments as blowing up arguments. The above ideas show the importance of studying problems on all of $\mathbb{R}^N$ or on half spaces even if we want to study bounded domain problems with small diffusion.

In a number of cases, the bounded domain problem (1) has many positive solutions when ϵ is small (cp. [14]). Thus is seems natural to consider a restricted class of problems, those which are in some sense stable or more generally "not too unstable". This is natural from the point of view of applications. Lastly for this introduction I should point out that problems such as (1) occur in many applications, for example in population models in biology, the theory of combustion and the theory of catalysts to name but a few. In many, but not all of these applications, ϵ is rather small and the solutions are positive. Note that, from the viewpoint of applications, it would also be natural to study systems. Note also that, for applications, the natural cases are $N = 2$ or $N = 3$.

1. Linearized Stable Solutions on $\mathbb{R}^N$

Here we consider bounded linearized stable solutions of

$$-\Delta u = f(u) \tag{3}$$

on $\mathbb{R}^N$, where u is said to be linearized stable if

$$J(\phi) \equiv \int (|\nabla\phi|^2 - f'(u)\phi^2) \geq 0 \text{ for all } \phi \in C_c^\infty(\mathbb{R}^N), \tag{4}$$

where f is assumed to be C^1, the intervals are over $\mathbb{R}^N$ and $C_c^\infty(\mathbb{R}^N)$ denotes the smooth functions of compact support in $\mathbb{R}^N$. In fact the choice of $C_c^\infty(\mathbb{R}^N)$ does not seem to be important. We could use any reasonable class of $W^{1,2}$ functions which decay reasonably rapidly at infinity.

Here there is a very natural and important conjecture. Assume $N \leq 8$ and u is a bounded linearized stable solutions of (3). Then either u is constant on $\mathbb{R}^N$ or, after a rotation of axis, $u = u(x_1)$. In the latter case, it is easy to deduce that $u = u(x_1)$ where u is monotone in x_1. It is easy to use the

first integral of the ordinary differential equation to completely classify these solutions. This conjecture is known to be true for $N = 2$ and is known to be false if $N \geq 11$ because there are non-radial linearized stable solutions for $f(y) = y^p$ for certain large p (cp. [17]). (In fact, as we see below, it fails for $N = 9$). Note that the above conjecture is a strengthening of the well known De-Georgi conjecture where our condition u is linearized stable is strengthened to u is strictly monotone in some direction. The above conjecture seems the more natural one for applications. Note that the De-Georgi conjecture is known for $N = 3$ by [1] and the counterexample in [15] shows that it fails for $N \geq 9$. Savin [21] has proved the De-Georgi conjecture for $4 \leq N \leq 8$ under additional conditions on f and on the behaviour of u as $x_1 \to \pm\infty$ (where u is strictly monotonic in x_1).

We now discuss our main conjecture.

Theorem 1.1. *If u is a bounded linearized stable solution of (3) and there is a $C > 0$ and a sequence $R_i \to \infty$ such that $\int_{B_{R_i}} |\nabla u|^2 \leq CR_i^2$, then u is constant or $u = u(x_1)$ after a rotation of axes.*

This appears in [9], though there are closely related results in [1] and the work of Gui-Ghoussoub. The idea is to first use the linearized stability to prove that there exists a positive function h on $\mathbb{R}^N$ and $\mu \geq 0$ such that

$$-\Delta h = f'(u)h + \mu h \tag{5}$$

on $\mathbb{R}^N$. Unfortunately, we have almost no control on h at infinity. Note that $\frac{\partial u}{\partial x_i}$ satisfies (5) with $\mu = 0$. We can use an old ordinary differential equation trick by obtaining an equation for $\frac{\partial u}{\partial x_i}/h$ (which gets rid of the $f'(u)$ terms) and we use a clever test function argument of Ambroseo and Cabre [2] to prove $\frac{\partial u}{\partial x_i}/h$ is constant. The result can be proved rather easily from this.

Theorem 1.1 has some nice applications. Firstly if u is a bounded solution of (3) on $\mathbb{R}^N$, then standard elliptic estimates imply ∇u is bounded on $\mathbb{R}^N$. Hence it follows from Theorem 1.1 that our main conjecture is true for $N = 2$. (Note that it is open if $3 \leq N \leq 8$.) Moreover it is true if $N \leq 4$ and f has fixed sign because in this case one can easily use test function arguments (or the divergence theorem) to prove the conditions of Theorem 1.1. (See [9] for $N = 3$ and [16] for $N = 4$.) Note that, usually, the main difficulty in proving the required inequality in Theorem 1.1 is to prove the corresponding inequality for $\int_{B_R} uf(u)$.

There is another method due to Farina [17] which tends to give better results if $N > 4$. First assume that $u \geq 0$ and $f(y) = y^p$ where $p > 1$. There is a restriction on p if $p \geq 11$. The idea is to multiply the equation for u by $\psi(u)(l(x))^2$ and substitute $\phi(x) = s(u)l(x)$ in the linearized stability equation where $\psi' = (s')^2$, $\psi(0) = 0$, $s(0) = 0$ and $l(x)$ is a smooth function of compact support. If we then subtract the two equations, we get an inequality involving no derivatives of u. (The idea of doing this in the bounded domain case seems to

go back to Crandall and Rabinowitz [6]). By a number of elementary but very clever tricks and, by careful choice of l, Farina proves the result. He can allow u to change sign if $f(y) = |y|^{p-1}y$. His result is essentially optimal for these nonlinearities. More recently, he and Dupaigne [16] obtained similar results if f has similar behaviour near zero, f is convex on $[0, \infty)$ and f is concave on $(-\infty, 0)$.

In fact it is easy to localize the idea somewhat. For simplicity assume that $N \leq 10$. If $p > 1$, $f(0) = 0$, $f'(y)/y^{p-1} \to a > 0$ as $y \to 0^+$ and $f'(y) > y^{-1}f(y)$ for $y > 0$, one can prove that bounded linearized stable solutions u on $\mathbb{R}^N$ satisfy $u \leq 0$ on $\mathbb{R}^N$. This is very useful because it settles our conjecture for $N \leq 4$ for non linearities f which are also of fixed sign for $y \leq 0$ by applying our earlier results. (One can do somewhat better for f's where 0 is a simple zero.)

There is a special case of our conjecture which should be easier but is very important for bounded domain problems. Assume u is a positive linearized stable solution of (3) on $\mathbb{R}^3$ such that $u \to 0$ as $x_1 \to -\infty$ uniformly in x_2, x_3. Then we conjecture $u = u(x_1)$. This is known if f has fixed sign or if f is not too flat at any of its zeros (cp. [9] and [10]). The idea is to try to use moving plane ideas to prove that u is monotone in x_1. (The result would then follow from this by [1]).

We close this section with a brief discussion of the half space case for the Dirichlet boundary condition pointing out the differences. Firstly, for positive bounded solutions, one can prove the only solutions are functions of x_1 if $N \leq 3$ or if the only linearized stable solutions of (3) on $\mathbb{R}^{N-1}$ are constants or functions of one variable (cp. [4] and related papers of theirs). If we also require that the solutions are linearized stable, we can replace "positive" by "non-negative". Secondly, if $N = 3$ and the only bounded linearized stable solutions of (3) on $\mathbb{R}^N$ are constant, one can prove under very weak additional assumptions that there are no bounded sign changing linearized stable solutions on the half space (for Dirichlet boundary conditions).

We have discussed the linearized stable solution on $\mathbb{R}^N$ in some detail because it is the basis for all the later sections.

2. Finite Morse Index Solutions on $\mathbb{R}^N$

A solution u of (3) is said to have finite Morse index if there is a subspace W of $C_c^\infty(\mathbb{R}^N)$ such that W has finite co-dimension in $C_c^\infty(\mathbb{R}^N)$ and such that $J(\phi) \geq 0$ on W. (Here W has finite co-dimension in $C_c^\infty(\mathbb{R}^N)$ means that the quotient space $C_c^\infty(\Omega)/W$ is finite dimensional. The co-dimension of W is the dimension of this quotient.) The Morse index is the co-dimension of W when a maximal W is chosen. Intuitively a finite Morse index solution is "not too unstable". If u is bounded, this is clearer because it is not difficult to prove that our condition is equivalent to proving that the unbounded self-adjoint operator $-\Delta - f'(u)I$ on $L^2(\mathbb{R}^N)$ has finitely many negative points in the spectrum each of which is an eigenvalue of finite multiplicity. However the above condition is

easier to work with. Note also, as before, the choice $C_c^\infty(\mathbb{R}^N)$ is not crucial. Note also that the condition that the Morse index is at most k behaves well under limits. This is crucial in many proofs.

For the rest of this section we assume the condition that the only bounded linearized stable solutions of (3) on $\mathbb{R}^N$ are constants. We call this condition (Z). Indeed, almost nothing is known about the finite Morse index solutions in other cases, even if $N = 2$. It seems much more complicated. This is an interesting open question.

Suppose that condition (Z) holds and the zeros of f are isolated. Then it is easy to prove that, if u is a bounded finite Morse index solution of (3) on $\mathbb{R}^N$, then there is a zero c of f such that $u(x) \to c$ as $||x|| \to \infty$ (where $f'(c) \leq 0$ if f is C^1).

If we strengthen the condition on f near zeros of f we can prove much more. We assume condition Z. In addition we assume condition (*): For each non-simple zero c of f, assume either that $f'(x) \leq 0$ in a neighbourhood of c or there exist $p > 1$, $q > p - 1$ and c_+, c_- both not zero such that $f'(x) \sim c_\pm |x - c|^{p-1} + O|x - c|^q$ for x close to c. If $N \geq 11$, we also need to assume that $p < p_c(N)$ where $p_c(N)$ is the Joseph-Lungren exponent (cp. [17]). Note that $p_c(N) = \infty$ if $N \leq 10$. Then we have the following theorem:

Theorem 2.1. *Assume that condition (*) and condition Z both hold and u is a bounded finite Morse index solution of (3) on $\mathbb{R}^N$. Then $\nabla u \in L^2(\mathbb{R}^N)$, $(u - c)f(u)$ and $F(u) \in L^1(\mathbb{R}^N)$ (where $F' = f$ and $F(c) = 0$) and*

$$\int [(N-2)(2N)^{-1}(u-c)f(u) - F(u)] = 0.$$

Moreover, if $u - c$ has fixed sign on $\mathbb{R}^N$, then u is radial (up to translations).

Remark 1. In applying the last part of the theorem it is useful to note that if u is a bounded solution of (3) on $\mathbb{R}^N$, then $f(\sup u) \geq 0$ and $f(\inf u) \leq 0$. This is useful for proving that $u - c$ has fixed sign.

Remark 2. The equality is known as the Pokojaev identity. The finite Morse index condition is used to prove that we have good enough decay to prove the Pokojaev identity.

Remark 3. It is an interesting open question to prove the radial result without assuming finite Morse index but assuming $u(x) \to c$ as $||x|| \to \infty$

Remark 4. Condition (Z) holds if $\pm f$ are strictly convex or f' is strictly convex provided condition (*) holds. This is closely related to the work of Gladiali, Pacella and Weth [18] who assume $\nabla u \in L^2(\mathbb{R}^N)$ but do not assume $u \in L^\infty(\mathbb{R}^N)$ and do not assume the condition (*). Simple examples show that the condition on p is necessary for Theorem 2.1. Note also that Gladiali et al. prove under their conditions that solutions of Morse index at most N are radial. The condition $u \in L^\infty(\mathbb{R}^N)$ rather than $\nabla u \in L^2(\mathbb{R}^N)$ seems much more convenient for applications. On the other hand examples in Bartsch-Willem [3] and Musso,

Pacard and Wei [20] show that for many f's with f' strictly convex there may exist non-radial finite Morse index solutions such that $u \to c$ as $||x|| \to \infty$ (even examples where $f'(c) < 0$).

The proof of Theorem 2.1 starts by using linearized stability results and scaling to obtain $||x||^{-2/(p-1)}$ decay of u and then using classical techniques to greatly improve the decay if p is not too small (in particular if $p > \frac{N+2}{N-2}$). The last part follows from the decay estimates and a moving plane argument. (A moving plane argument is a geometric argument which compares $u(x)$ with $u(x^\lambda)$ where x^λ is the reflection of x in some hyperplane.) It seems much harder to understand the finite Morse index solutions on $\mathbb{R}^N$ in the case where there is a 1-dimensional bounded strictly monotone solution of (3).

3. Application to Bounded Domain Problems

Here we are interested in applications of (1) where ϵ is small or we are interested in large solutions. We restrict ourselves to the Dirichlet problem, though our technique could be used for other boundary conditions. We first consider the former case:

Theorem 3.1. *Assume that $K > 0$, f is C^1, f has isolated zeros, and that (I) $N = 2$ or (II) condition Z holds or (III) $N = 3$ and a weak technical condition holds near the non-simple zeros of f. Then the non-trivial stable positive solutions of (1) with $\sup u \leq K$ are close to a positive zero C of f in the interior of Ω and near $\partial\Omega$ are close after rescaling to a positive solution of*

$$-v''(t) = f(v(t)) \tag{6}$$

$v(0) = 0$, $v > 0$ on $(0, \infty)$, $v \to C$ as $t \to \infty$. There is exactly one positive stable solution for each C such that (6) has a positive solution. In particular the number of stable positive solutions is independent of the shape of Ω.

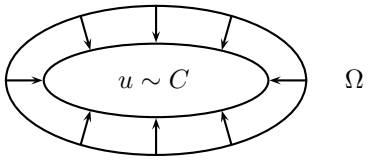


Figure 1.

The proof of this is in two parts. Firstly, we use the results of the previous section for $\mathbb{R}^N$ and half spaces and limit arguments to prove that any stable solution has the above form if ϵ is small. Note that in some cases it is easier to start from the interior while in other cases it is easier to start from the

boundary and move into the interior. To prove the converse one uses more classical techniques such as sub and supersolutions and blow ups.

If condition Z holds, it turns out that there are no stable changing sign solutions. However if there is a non-constant monotone bounded solution of $-y'' = f(y(t))$ on $\mathbb{R}$, there may sometimes be changing sign stable solutions (cp. [19]). This uses gamma-convergence ideas. Whether these occur depends on the shape of Ω. This case is poorly understood even for $N = 2$.

Suppose that condition Z holds and the zeros of f are nodal. Then it is possible to use similar ideas to prove that when $k > 0$, then for small ϵ, positive solutions of (1) of Morse index at most k are asymptotically of the form $\phi_\epsilon(x) + \sum_{i=1}^{s}(W_i(\epsilon^{-1}(x - x_i) - C)$ $1 \leq s \leq k$, where ϕ_ϵ is a positive stable solution of (1) near C on most of Ω (or identically zero), W_i is a solution of (3) on $\mathbb{R}^N$ such that $W_i > C$ on $\mathbb{R}^N$, and $W_i \to C$ as $|x| \to \infty$ and $x_i \in \Omega$. In other words, positive finite Morse index solutions look like a stable solution with a finite number of sharp peaks superimposed. Under weak conditions we can prove that the W_i are radial.

If $N = 3$, condition Z can be largely removed though it is unclear if it can be completely removed. If we allow u to change sign, we obtain a similar result except that we delete the requirement that $W_i > C$ and W_i need not be radial. The situation is much less clear if $N = 2$ and there is a non-constant bounded monotone solution of $-y'' = f(y)$ on $\mathbb{R}$. Also the location of the peaks are not completely understood. (If $C = 0$, and we only have one peak, we have a rather complete understanding.)

If we allow non nodal zeros, the theory is essentially the same except that we may sometimes have unstable finite Morse index solutions which look to first order like stable solutions. In the interior of the domain these solutions rescaled look like a changing sign solution of

$$\begin{aligned} -\Delta v =& |v|^p \quad \text{in } \Omega \\ & v \to \infty \text{ as } x \to \partial\Omega. \end{aligned} \tag{7}$$

Here p is the order of the zero C of f. This is discussed in [12]. This shows a surprising connection with problems with infinite boundary values.

Lastly, our methods can be used to help prove that the branch of positive solutions of (1) bifurcating from $(0, 0)$ (where we use u and $\lambda = \epsilon^{-2}$ as variables) has infinitely many bifurcations if f is real analytic, if $f(y) > 0$ for $y > 0$, if $f(0) > 0$, if $f'(y) \sim e^y$ as $y \to \infty$ and if $3 \leq N \leq 9$. (If $f(0) = 0$), $f'(0) > 0$, the result is still true though the branch may not bifurcate from $(0, 0)$.) Note that the restriction on N is sharp even if Ω is a ball. There is a similar theorem if $f'(y) \sim y^{p-1}$ as $y \to \infty$ if $3 \leq N \leq 10$, p is large and Ω is star-shaped. (We do not know if the star-shapedness condition can be removed.)

The idea here is to use real analytic bifurcation theory (see [5] and [6]) to find an unbounded arc A of positive solutions bifurcating from $(0, 0)$ such that the linearisation (in u) of the equation is invertible except at isolated points of the arc. This is where we use the real analyticity. (By an unbounded arc

we mean a homeomorph of $[0,1)$.) The above result need not be true for C^∞ maps.) By obtaining a limit equation on $\mathbb{R}^N$, we prove that the Morse index of a solution (u,λ) tends to infinity as $||u||_\infty + |\lambda| \to \infty$ (The key result here is to prove (see [11] and [13]) that the limiting problem $-\Delta u = e^u$ on $\mathbb{R}^N$ has no negative finite Morse index solution. This is where the condition $3 \leq N \leq 9$ is used.) Thus there must be infinitely many points of A where the Morse index changes. Finally, we use local topological invariants such as critical groups or Conley indexes to prove that any point where the Morse index changes is a bifurcation point. This uses the variational structure of the equation. Lastly, the real analyticity can be avoided but the interpretation is then not so clear. Moreover, the results can be proved for other asymptotic behaviours at infinity.

Lastly, a general question is to obtain similar theories for interesting general classes of elliptic systems.

References

[1] Alberti, Giovanni and Ambrosio, Luigi and Cabré, Xavier, *On a long-standing conjecture of E. De Giorgi: symmetry in 3D for general nonlinearities and a local minimality property, Special issue dedicated to Antonio Avantaggiati on the occasion of his 70th birthday*, Acta Appl. Math., **65**, (2001), 9–33.

[2] Ambrosio, Luigi and Cabré, Xavier, *Entire solutions of semilinear elliptic equations in* $\mathbb{R}^3$ *and a conjecture of De Giorgi*, J. Amer. Math. Soc. **13** (2000),725–739.

[3] Bartsch, Thomas and Willem, Michel, *Infinitely many nonradial solutions of a Euclidean scalar field equation*, J. Funct. Anal. **117**, (1993), 447–460.

[4] Berestycki, Henri and Caffarelli, Luis and Nirenberg, Louis, *Further qualitative properties for elliptic equations in unbounded domains*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), **25**, (1997), 69–94.

[5] Buffoni, Boris and Toland, John, *Analytic theory of global bifurcation*, Princeton Series in Applied Mathematics, Princeton University Press, (2003), x+169.

[6] Crandall, Michael G. and Rabinowitz, Paul H. *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Rational Mech. Anal. **58**, (1975), 207–218.

[7] Dancer, E. N. *Global structure of the solutions of non-linear real analytic eigenvalue problems*, Proc. London Math. Soc. (3), **27**, (1973), 747–765.

[8] Dancer, E. N. *Stable and finite Morse index solutions on* $\mathbb{R}^n$ *or on bounded domains with small diffusion. II*, Indiana Univ. Math. J. **53**, (2004), 97–108.

[9] Dancer, E. N. *Stable and finite Morse index solutions on* $\mathbb{R}^n$ *or on bounded domains with small diffusion*, Trans. Amer. Math. Soc. **357**, (2005), 1225–1243 (electronic).

[10] Dancer, E. N. *Finite Morse index solutions of supercritical problems*, J. Reine Angew. Math. **620**, (2008), 213–233.

[11] Dancer, E. N. *Finite Morse index solutions of exponential problems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, **25**, (2008), 173–179.

[12] Dancer, E. N. *Stable and finite Morse index solutions for Dirichlet problems with small diffusion in a degenerate case and problems with infinite boundary values*, Adv. Nonlinear Stud. **9**, (2009), 657–678.

[13] Dancer, E. N. and Farina, Alberto, *On the classification of solutions of* $-\Delta u = e^u$ *on* $\mathbb{R}^N$*: stability outside a compact set and applications*, Proc. Amer. Math. Soc. **137**, (2009), 1333–1338.

[14] Dancer, E. N. and Yan, Shusen, *A singularly perturbed elliptic problem in bounded domains with nontrivial topology*, Adv. Differential Equations, **4**, (1999), 347–368.

[15] Del Pino, W. and Kowalczyk, M. and Wei, J. *On De Georgi's conjecture in dimension* $N \geq 9$, preprint.

[16] Dupaigne, Louis and Farina, Alberto, *Stable solutions of* $-\Delta u = f(u)$ *in* $\mathbb{R}^N$, preprint.

[17] Farina, Alberto, *On the classification of solutions of the Lane-Emden equation on unbounded domains of* $\mathbb{R}^N$, J. Math. Pures Appl. (9), **87**, (2007), 537–561.

[18] Gladiali, F. and Pacella, F. and Weth, T. *Symmetry and nonexistence of low Morse index solutions in unbounded domains.*, J. Math. Pure. Appl. to appear.

[19] Kohn, Robert V. and Sternberg, Peter, *Local minimisers and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, **111**, (1989), 69–84.

[20] Musso, M. and Pacard, F. and Wei, J. *Finite energy, sign changing solutions with dihedral symmetry for stationary non linear Schrodinger equations*, preprint.

[21] Savin, Ovidiu, *Regularity of flat level sets in phase transitions*, Ann. of Math. (2), **169**, (2009), 41–78.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Almgren's Q-valued Functions Revisited

Camillo De Lellis*

Abstract

In a pioneering work written 30 years ago, Almgren developed a far-reaching regularity theory for area-minimizing currents in codimension higher than 1. Building upon Almgren's work, Chang proved later the optimal regularity statement for 2-dimensional currents. In some recent papers the author, in collaboration with Emanuele Spadaro, has simplified and extended some results of Almgren's theory, most notably the ones concerning Dir-minimizing multiple valued functions and the approximation of area-minimizing currents with small cylindrical excess. In this talk I will give an overview of our contributions and illustrate some possible future directions.

Mathematics Subject Classification (2010). Primary 49Q20; Secondary 35J55, 54E40, 53A10 .

Keywords. Area-minimizing currents , regularity theory , multiple-valued functions, analysis on metric spaces, higher integrability.

1. Introduction

1.1. The regularity theory for area-minimizing currents.

In this note we will describe some recent contributions to the regularity theory for integer rectifiable area-minimizing currents. For the sake of simplicity we will restrict ourselves to currents in the Euclidean space. For all the relevant definitions concerning currents we refer the reader to the classical textbooks [16] and [39].

As it is well known there is a dramatic difference in the theory depending on the codimension of the current. In codimension 1 currents without boundary

*Camillo De Lellis, Insitut für Mathematik, Universität Zürich, Zürich, Switzerland.
E-mail: camillo.delellis@math.uzh.ch.

are boundaries of sets of finite perimeter. This allows several important simplifications in the theory (see for instance [23]) and it also implies that area-minimizing currents of codimension 1 enjoy much better regularity properties. Let us briefly review the main results in the interior regularity theory.

Codimension 1. Let T be an area-minimizing current of dimension n in $\mathbb{R}^{n+1}$.

(a1) For $n \leq 6$, T is an analytic submanifold in $\mathbb{R}^{n+1} \setminus \operatorname{supp}(\partial T)$ (see for instance [16, Theorem 5.4.15]);

(a2) for $n = 7$, T is an analytic submanifold in $\mathbb{R}^{n+1} \setminus \operatorname{supp}(\partial T)$ with the exception of a discrete set $\operatorname{Sing}(T)$ of singular points (see for instance [16, Section 5.4.16]);

(a3) for $n = 7$, in a neighborhood of each $x \in \operatorname{Sing}(T)$ the current is a perturbation of an area-minimizing cone (see [40]);

(a4) for $n > 7$, T is an analytic submanifold in $\mathbb{R}^{n+1} \setminus \operatorname{supp}(\partial T)$ with the exception of a closed set $\operatorname{Sing}(T)$ of (Hausdorff) dimension at most $n - 7$ (see for instance [39, Theorem 37.7]);

(a5) if $n > 7$, the singular set $\operatorname{Sing}(T)$ is rectifiable and has locally finite $\mathcal{H}^{n-7}$-measure (see [42, Lecture 4, Theorem 4] and [41]; here $\mathcal{H}^{\alpha}$ denotes, as usual, the α-dimensional Hausdorff measure).

The results in (a2) and (a5) give the optimal estimates of the size of $\operatorname{Sing}(T)$. The optimality of (a2) is shown by the Simons cone. The minimizing property of this cone was first proved in the celebrated paper of Bombieri, De Giorgi, and Giusti [8]. In order to prove the optimality of (a5) it suffices to take the product of the Simons cone with a linear space of dimension $n - 7$ (cp. with [16, Theorem 5.4.9]).

Codimension $k > 1$. Let T be an integer rectifiable area-minimizing current of dimension n in $\mathbb{R}^{n+k}$.

(b1) If $n = 1$, T is the union of nonintersecting straight lines;

(b2) if $n = 2$, T is an analytic submanifold in $\mathbb{R}^{n+k} \setminus \operatorname{supp}(\partial T)$ with the exception of a discrete set $\operatorname{Sing}(T)$ (see [7]);

(b3) if $n = 2$, in a neighborhood of each $x \in \operatorname{Sing}(T)$ the current is a perturbation of a suitable "branched holomorphic curve" (see [32]);

(b4) for $n > 2$, T is an analytic submanifold in $\mathbb{R}^{n+k} \setminus \operatorname{supp}(\partial T)$ with the exception of a closed set $\operatorname{Sing}(T)$ of dimension at most $n - 2$ (see [4]).

The size estimate of (b2) is optimal, as shown by taking any holomorphic curve in $\mathbb{R}^4 = \mathbb{C}^2$ with branch points. This example plays a crucial role in the rest of our discussion and will be examined in further detail later on.

One first striking difference between these series of results is that in the latter singularities appear quite naturally as soon as we depart from the trivial case $n = 1$. Moreover, this appearance, linked to the well-known phenomenon of branching of holomorphic curves, is far much easier to understand than the minimizing property of the Simons cone, which is the simplest example of a singular area-minimizing current with codimension 1.

The second striking difference is in the length, the intricacy and the technical complications presented by Almgren's and Chang's results ((b2) and (b4)) in comparison with Federer's size estimates of $\mathrm{Sing}(T)$ ((a2) and (a4)). Assuming indeed a certain amount of prerequisites in geometric measure theory, (a1), (a2) and (a4) are essentially the combination of three ingredients: the pioneering work of De Giorgi on the excess-decay [9], the classical work of Simons on stable minimal cones [43] and Federer's reduction argument, see [17]. Moreover, only a relatively small portion of the theorems in [43] are needed to prove (a4). Let me also mention that, before the work of [43] completed the proof of (a1), lower-dimensional versions were achieved in the works of Fleming, De Giorgi and Almgren [19, 10, 3]. The interested reader might find a complete and quite readable account in the beautiful book of Giusti [23].

Assuming the same amount of prerequisites, the theorem in (b4) is instead a monograph of about 950 pages, see [4]. This monograph contains, among many other things, far-reaching generalizations of both De Giorgi's and Federer's arguments. The proof of (b2) is contained in the paper [7], where the author builds upon (essentially all) the techniques developed in Almgren's monograph and on the important papers [33] and [47]. Indeed, some of the constructions needed in [7] are claimed to be suitable modifications of the ones in [4], but the detailed proofs of these statements have never appeared.

1.2. Branching. Let us examine in more details the first obstruction to the full regularity in the case of higher codimension. The key observation relies on a classical computation of Wirtinger [49], used by Federer in his elegant proof of the following statement (cp. to [16, Section 5.4.19]).

Theorem 1.1. *If M is a Kähler manifold of real dimension $2m$ and Γ a complex submanifold of M of real dimension $2j$, then Γ represents an integer rectifiable area minimizing current. More precisely, if U is a bounded open set with $U \cap \mathrm{supp}\,(\partial\Gamma) = \emptyset$ and Σ is an integer rectifiable current of dimension $2j$ such that*

- $\partial(\Gamma - \Sigma) = 0$,
- $\mathrm{supp}\,(\Gamma - \Sigma) \subset U$,

then the mass of Σ in U is larger than the mass of Γ in U. Moreover, the inequality is strict unless $\Gamma = \Sigma$.

In a more modern language, the Wirtinger-Federer result can be rephrased in the following way: the k-th exterior power of the Kähler form is a calibration

for holomorphic submanifolds of complex dimension k. For a beautiful account of calibrating forms we refer the reader to the paper [27].

The presence of branching phenomena in area-minimizing currents of codimension larger than 1 is also the principal reason for the difficulty of Almgren's monumental result. Much of Section 2 will be devoted to give an intuitive explanation of this.

1.3. Looking for a manageable proof. The intricacy of Almgren's big regularity paper [4] has essentially stopped the research in the area till few years ago, in spite of the abundance of interesting geometric objects which are naturally minimal submanifolds of "large" codimension (see again the paper [27]). Recently, in view of some applications to geometry and topology, alternative proofs of Chang's result have been found for J-holomorphic curves. The first of these proofs has been given by Taubes in [45] for J-holomorphic curves in symplectic 4-manifolds. The generalization of Taubes' approach to $1-1$ currents in (even-dimensional) manifolds carrying a certain complex structure has been given by Rivière and Tian (see [36], [35] and [37]). This proof contains several beautiful ideas and faces some of the same problems which are solved in Almgren's monograph. However, its applicability seems limited to 2-dimensional currents which are calibrated by some complex structure. At present, the general theorem of Chang (not to speak of the result of Almgren) does not seem reachable with similar approaches.

The remarkable papers [35] and [37] and several discussions of the author with Tristan Rivière have been the starting point of the line of research which will be presented here. The results which will be described in this note have appeared in the papers [13], [14], [12], [44] and [15]. A substantial part of these papers is dedicated to give self-contained and much simpler proofs of a considerable portion of Almgren's monograph. In the remaining part we take advantage of some new ideas to expand Almgren's theories in other directions. Though some fundamental ideas behind these papers are still the ones of Almgren, our approaches highlight some rather new aspects. In some cases we have taken advantage of modern techniques of metric analysis, in some other we have discovered new phenomena. The overall result is that we can handle the complexity of the subject in a much more efficient way. Our obvious final goal is to give a less complex, yet complete account of Almgren's and Chang's regularity results and possibly go beyond them in a not so far future.

In the next sections we will describe roughly the contents of the papers [13], [14] and [15]. In the final section we collect several interesting related open problems.

2. Why Multiple Valued Functions?

2.1. De Giorgi's excess decay. The first breakthrough of the regularity theory for area-minimizing currents is due to De Giorgi. In order to state

De Giorgi's main theorem, we have to introduce the so-called (spherical) excess $\mathrm{Ex}(T, B_r(p))$ of the current T in the ball $B_r(p)$. For every simple unitary n-vector $\vec{\pi}$, we set

$$\mathrm{Ex}(T, B_r(p), \pi) := \frac{1}{2} \fint_{B_r(p)} |\vec{T} - \vec{\pi}|^2 d\|T\| \,. \tag{2.1}$$

The measure $\|T\|$ is the localized mass of the current: for every open set U, $\|T\|(U)$ is the total mass of the current in U. $\vec{T}$ is the simple unitary n-vector field orienting T.

The spherical excess is then defined as

$$\mathrm{Ex}(T, B_r(p)) \;:=\; \min_{\pi} \mathrm{Ex}(T, B_r(p), \pi) \,.$$

This definition is valid in any codimension. For the reader who is not very familiar with the notation of geometric measure theory, the formulas can be considerably simplified in codimension 1. First of all, the minimum can be taken over all oriented n-dimensional planes π ($\vec{\pi}$ is then just the unitary n-vector orienting π). Moreover $|\vec{T} - \vec{\pi}|$ can be substituted by $|\nu_T - \nu|$, where:

- ν_T is the unit vector field normal to the current, compatible with the orientation of the tangent n-vector $\vec{T}$;
- ν is the unit vector normal to π compatible with the orientation $\vec{\pi}$.

A third important object that we need to introduce is the density of the current at a point, which is defined as

$$\theta(T, p) \;:=\; \lim_{r \downarrow 0} \frac{\|T\|(B_r(p))}{\omega_n r^n} \,, \tag{2.2}$$

where ω_n denotes, as usual, the n-dimensional measure of the n-dimensional ball. The existence of the limit in (2.2) is guaranteed by the monotonicity formula (cp. with [39, Section 4.17]).

Theorem 2.1. *Let Q be a positive integer. There exist constants $\varepsilon, \beta > 0$ depending only on Q and n such that the following holds. Let T be an area-minimizing integral current of dimension n in $\mathbb{R}^{n+1}$. Assume that, for $r > 0$ and $p \in \mathrm{supp}\,(T)$, the following hypotheses are satisfied:*

(i) $\theta(T, p) = Q$;

(ii) $\mathrm{supp}\,(\partial T) \cap B_r(p) = \emptyset$;

(iii) $\|T\|(B_r(p)) \leq (Q + \varepsilon)\omega_n r^n$;

(iv) the spherical excess of T in $B_r(p)$ is smaller than ε.

Then $\mathrm{supp}\,(T) \cap B_{r/2}(p)$ is the graph of a $C^{1,\beta}$ function f.

To be more precise, De Giorgi in [9] proved the case $Q = 1$ of this theorem. However the general case $Q > 1$ can be easily recovered from De Giorgi's statement using the decomposition of T in boundaries of sets of finite perimeter as in [16, Section 4.5.17].

To get some intuitive idea about the theorem above, consider the extreme case where the spherical excess in $B_r(p)$ is 0. Using assumption (ii) we then conclude that T in $B_r(p)$ consists of (possibly countably many) parallel disks. Exploiting (i), (iii) and the minimality of T, from the monotonicity formula we easily conclude that, in a slightly smaller ball $B_{r-C\varepsilon}(p)$, T consists of a single disk containing the origin and counted with multiplicity Q. Thus, the assumptions (i)–(iv) tell us that the current T is close, in an "average" sense, to Q copies of a single disk. Theorem 2.1 could be therefore classified as an "ε-regularity theorem".

2.2. Again branching. As already mentioned, De Giorgi's original proof covers the case $Q = 1$ and the extension to $Q > 1$ uses heavily the features of codimension 1 currents. In higher codimension the statement is still correct for $Q = 1$ (see for instance [16, Theorem 5.4.7]; in fact much more is true, see [2]), but fails dramatically if $Q > 1$. Once again, the main reason for this breakdown is the existence of branching points.

Remark 2.2. Consider in $\mathbb{R}^4 = \mathbb{C}^2$ the holomorphic curve $\Gamma = \{(z, w) : z^2 = w^3\}$. Theorem 1.1 implies that Γ is an area–minimizing current of real dimension 2 in any bounded open subset of $\mathbb{R}^4$. Moreover, set $p = 0$. Then

$$\theta(T, 0) = \lim_{r \downarrow 0} \frac{\|T\|(B_r(0))}{\omega_2 r^2} \;=\; 2\,.$$

Obviously, given any positive $\varepsilon > 0$ there is a δ such that (i)–(iv) are satisfied for every $r < \delta$. On the other hand, no matter how small r is, $B_r(0) \cap \Gamma$ is never the graph of a smooth function.

We proceed our discussion by giving an oversimplified description of De Giorgi's proof of Theorem 2.1 in the case $Q = 1$. In a first step, the hypotheses (i)–(iv) are used to approximate the current T with the graph G of a Lipschitz (real valued) function f with small Lipschitz constant. In particular, the approximation algorithm ensures that the area of T and the area of G are close. On the other hand, recall that the area of the graph of a function over a domain Ω is given by the formula

$$\int_\Omega \sqrt{1 + |\nabla f|^2}\,. \tag{2.3}$$

If $|\nabla f|$ is small, this integral is close to

$$\int_\Omega \left(1 + \frac{|\nabla f|^2}{2}\right) \tag{2.4}$$

(in higher codimension, i.e. when f is vector-valued, the formula for (2.3) is more complicated, but the second order expansion is nonetheless given by (2.4)).

Thus, the minimality of the current T implies that f is close, in a suitable integral sense, to a minimum of the Dirichlet energy, i.e. to an harmonic function. Using the decay properties of harmonic functions, one can infer that the excess $\mathrm{Ex}(T, B_\rho(p))$ is decaying like $\rho^{2\beta}$ for some $\beta > 0$. This decay leads then to the $C^{1,\beta}$ regularity via a "Morrey-type" argument.

2.3. Dealing with branching. As already noticed, in codimension 1 the higher multiplicity case can be reduced to the case of multiplicity 1. Obviously, Remark 2.2 shows that this reduction is impossible in codimension larger than 1. In that example the very beginning of De Giorgi's strategy fails, since it is simply not possible to approximate efficiently Γ with the graph of a (single valued) function. This discussion motivates the starting idea of Almgren's monograph. In order to tackle the regularity question in codimension larger than 1 we need to approximate currents with "multiple valued functions".

It is interesting to notice that, if we turn our attention to stationary currents (or, more generally, stationary integral varifolds), the reduction to multiplicity 1 becomes false even in the codimension 1 case. In this setting, the best result available at present is Allard's Theorem [2], which ensures regularity in a dense open set. Nothing better is known, even assuming stability, in spite of the fact that all available examples have singularities of dimension at most $n-1$. If we assume stability and an a-priori knowledge that the singular set has zero $\mathcal{H}^{n-2}$-measure, then the classical curvature estimates of Schoen and Simon imply that the singular set has in fact dimension at most $n-7$ (see [38]). In a very recent paper [48], Wickramasekera has extended this result to the optimal assumption that the $\mathcal{H}^{n-1}$-measure of the singular set is 0. Related questions are open for "stationary multiple valued functions" as well (see Section 8 below).

3. The Dirichlet Energy for Multiple Valued Functions

3.1. The metric space of unordered Q-tuples. Roughly the first fifth of Almgren's monograph is devoted to develop the theory of multiple valued functions. The obvious model case to keep in mind is the following. Given two integers k, Q with $\mathrm{MCD}(k, Q) = 1$, look at the function which maps each point $z \in \mathbb{C}$ into the set $M(z) := \{w^k : w^Q = z\} \subset \mathbb{C}$. Obviously for each z we can order the elements of the set $M(z)$ as $\{u_1, \ldots, u_Q\}$. However, it is not possible to do it *globally* in such a way that the maps $z \mapsto u_i(z)$ are continuous.

This motivates the following definition. Given an integer Q we define a Q-valued map from a set $E \subset \mathbb{R}^m$ into $\mathbb{R}^n$ as a function which to each point $x \in E$ associates an unordered Q-tuple of vectors in $\mathbb{R}^n$. There is a fairly efficient formulation of this definition which will play a pivotal role in our discussion.

Following Almgren, we consider the group $\mathscr{P}_Q$ of permutations of Q elements and we let $\mathcal{A}_Q(\mathbb{R}^n)$ be the set $(\mathbb{R}^n)^Q$ modulo the equivalence relation

$$(v_1, \ldots, v_Q) \equiv (v_{\pi(1)}, \ldots, v_{\pi(Q)}) \qquad \forall \pi \in \mathscr{P}_Q \, .$$

The set $\mathcal{A}_Q(\mathbb{R}^n)$ can be naturally identified with a subset of the set of measures (cp. with [4] and [13, Definition 0.1]).

Definition 3.1 (Unordered Q-tuples). Denote by $[\![P_i]\!]$ the Dirac mass in $P_i \in \mathbb{R}^n$. Then,

$$\mathcal{A}_Q(\mathbb{R}^n) := \left\{ \sum_{i=1}^{Q} [\![P_i]\!] \, : \, P_i \in \mathbb{R}^n \text{ for every } i = 1, \ldots, Q \right\} .$$

This set has a natural metric structure; cp. with [4] and [13, Definition 0.2] (the experts will recognize the well-known Wasserstein 2-distance, cp. with [46]).

Definition 3.2. For every $T_1, T_2 \in \mathcal{A}_Q(\mathbb{R}^n)$, with $T_1 = \sum_i [\![P_i]\!]$ and $T_2 = \sum_i [\![S_i]\!]$, we set

$$\mathcal{G}(T_1, T_2) \, := \, \min_{\sigma \in \mathscr{P}_Q} \sqrt{\sum_i \left| P_i - S_{\sigma(i)} \right|^2} . \tag{3.1}$$

3.2. Almgren's extrinsic maps. The metric $\mathcal{G}$ is "locally euclidean" at most of the points. Consider for instance the model case $Q = 2$ and a point $P = [\![P_1]\!] + [\![P_2]\!]$ with $P_1 \neq P_2$. Then, obviously, in a sufficiently small neighborhood of P, the metric space $\mathcal{A}_2(\mathbb{R}^n)$ is isomorphic to the Euclidean space $\mathbb{R}^{2n}$. This fails instead in any neighborhood of a point of type $P = 2\,[\![P_1]\!]$. On the other hand, if we restrict our attention to the closed subset $\{2\,[\![X]\!] : X \in \mathbb{R}^n\}$, we obtain the metric structure of $\mathbb{R}^n$. A remarkable observation of Almgren is that $\mathcal{A}_Q(\mathbb{R}^n)$ is biLipschitz equivalent to a deformation retract of the Euclidean space (cp. with [4, Section 1.3]). For a simple presentation of this fact we refer the reader to [13, Section 2.1].

Theorem 3.3. *There exists $N = N(Q, n)$ and an injective $\boldsymbol{\xi} : \mathcal{A}_Q(\mathbb{R}^n) \to \mathbb{R}^N$ such that:*

(*i*) $\mathrm{Lip}(\boldsymbol{\xi}) \leq 1$;

(*ii*) *if $\mathcal{Q} = \boldsymbol{\xi}(\mathcal{A}_Q)$, then $\mathrm{Lip}(\boldsymbol{\xi}^{-1}|_{\mathcal{Q}}) \leq C(n, Q)$.*

Moreover there exists a Lipschitz map $\boldsymbol{\rho} : \mathbb{R}^N \to \mathcal{Q}$ which is the identity on $\mathcal{Q}$.

In fact much more can be said: the set $\mathcal{Q}$ is a cone and a polytope. On each separate face of the polytope the metric structure induced by $\mathcal{G}$ is euclidean, essentially for the reasons outlined a few paragraphs above (cp. again with [4, Section 1.3] or with [14, Section 6.1]).

3.3. The generalized Dirichlet energy. Using the metric structure on $\mathcal{A}_Q(\mathbb{R}^n)$ one defines obviously measurable, Lipschitz and Hölder maps from subsets of $\mathbb{R}^m$ into $\mathcal{A}_Q(\mathbb{R}^n)$. However, if we want to approximate area-minimizing currents with multiple valued functions and "linearize" the area functional in the spirit of De Giorgi, we need to define a suitable concept of Dirichlet energy. We will now show how this can be done naturally. However, the approach outlined below is not the one of Almgren.

Consider again the model case of $Q = 2$ and assume $u : \Omega \to \mathcal{A}_2(\mathbb{R}^n)$ is a Lipschitz map. If, at some point x, $u(x) = \llbracket P_1 \rrbracket + \llbracket P_2 \rrbracket$ is "genuinely 2-valued", i.e. $P_1 \neq P_2$, then there exist obviously a ball $B_r(x) \subset \Omega$ and two Lipschitz functions $u_1, u_2 : B_r(x) \to \mathbb{R}^n$ such that $u(y) = \llbracket u_1(y) \rrbracket + \llbracket u_2(y) \rrbracket$ for every $y \in B_r(x)$ (in this and similar situations, we will then say that there is a *regular selection* for u in $B_r(x)$, cp. with [13, Definition 1.1]). For each separate function u_i, the classical Theorem of Rademacher ensures the differentiability almost everywhere.

Recall that our ultimate goal is to define the Dirichlet energy so that it is a suitable approximation of the area of the graph of u. The "graph of u over $B_r(x)$" is simply to union of the graphs of the two functions u_i. When the gradients ∇u_i are close to 0, the area of each graph is close to

$$\int_{B_r(x)} \left(1 + \frac{1}{2}|\nabla u_i|^2\right) .$$

Thus, the only suitable definition of Dirichlet energy of u on the domain $B_r(x)$ is given by

$$\int_{B_r(x)} |Du|^2 \;:=\; \int_{B_r(x)} (|Du_1|^2 + |Du_2|^2) .$$

By an obvious localization procedure, this definition can be extended to the (open!) set $\Omega_2 \subset \Omega$ where u is genuinely 2-valued.

For each element z in the complement set $\Omega_1 := \Omega \setminus \Omega_2$, $u(z)$ is a single point counted with multiplicity 2. Then there is a Lipschitz map $v : \Omega_1 \to \mathbb{R}^n$ such that $u(z) = 2\,\llbracket v(z) \rrbracket$ for every $z \in \Omega_1$. Again in view of our goal, the only suitable definition of the Dirichlet energy of u over Ω_1 is twice the Dirichlet energy of v. We thus are left with only one possibility for the Dirichlet energy on the global set Ω:

$$\mathrm{Dir}(u, \Omega) \;:=\; \int_{\Omega_2} (|Du_1|^2 + |Du_2|^2) + 2 \int_{\Omega_1} |Dv|^2 .$$

This analysis can be obviously generalized to any positive integer Q, leading to a general definition of Dirichlet energy for Lipschitz multiple valued functions. The graphs of Lipschitz multiple valued functions carry naturally a structure of integer rectifiable currents (see [4, Section 1.6] or [14, Appendix C]). It is not difficult to see that, when the Lipschitz constant is small, the Dirichlet energy defined in this section is the second order approximation of the area of the corresponding graph (we refer the reader to [14, Section 2.3]).

Almgren's definition of Dir goes instead through a suitable concept of differentiability for multiple valued functions and a corresponding Rademacher's theorem (in [4] the derivation of this result is quite involved and a much simpler proof has been published in [24]). The arguments in [13, Section 1] easily show that the two points of view are equivalent. In fact the "stratification" strategy outlined above yields a fairly straightforward proof of Almgren's generalized Rademacher's Theorem (see [13, Section 1.3.2]).

Having established the correct notion of Dirichlet energy for Lipschitz functions, one could define the Sobolev space $W^{1,2}(\Omega, \mathcal{A}_Q(\mathbb{R}^n))$ through a "completion strategy": a measurable map $v : \Omega \to \mathcal{A}_Q(\mathbb{R}^n)$ is in $W^{1,2}$ if and only if there is a sequence of Lipschitz maps u_k converging to v a.e. and enjoying a uniform bound $\mathrm{Dir}(\Omega, u_k) \leq C$. The Dirichlet energy of v is then defined via a "relaxation procedure": $\mathrm{Dir}(\Omega, v)$ is the infimum of all constants C for which there is a sequence with the properties above.

Almgren's approach is again rather different. $W^{1,2}$ maps are defined as those maps u for which $\boldsymbol{\xi} \circ u$ is $W^{1,2}$. The Dirichlet energy is again defined via a suitable notion of approximate differentiability. In our paper [13] we start from a third definition of Dirichlet energy and Sobolev space. However, all these points of view are completely equivalent, as one can easily conclude from the arguments in [13, Section 4] (cp. in particular with the Lipschitz approximation technique of [13, Proposition 4.4]).

3.4. The cornerstones of the theory of Dir-minimizers.

We are now ready to state the three main theorems of Almgren concerning Dir-minimizers. Their proofs occupy essentially Chapters 1 and 2, i.e. the first fifth of Almgren's monograph. In what follows, Ω is always assumed to be a bounded open set with a sufficiently regular boundary (in fact, in order to give a complete account, we should have defined the trace at $\partial\Omega$ of $W^{1,2}$ multiple valued functions; we have avoided to enter in the details to keep our presentation short: the interested reader can consult, for instance, [13, Definition 0.7]).

Theorem 3.4 (Existence for the Dirichlet Problem). *Let $g \in W^{1,2}(\Omega; \mathcal{A}_Q)$. Then there exists a* Dir*-minimizing $f \in W^{1,2}(\Omega; \mathcal{A}_Q)$ such that $f|_{\partial\Omega} = g|_{\partial\Omega}$.*

Theorem 3.5 (Hölder regularity). *There is a constant $\alpha = \alpha(m, Q) > 0$ with the following property. If $f \in W^{1,2}(\Omega; \mathcal{A}_Q)$ is* Dir*-minimizing, then $f \in C^{0,\alpha}(\Omega')$ for every $\Omega' \subset\subset \Omega \subset \mathbb{R}^m$. For two-dimensional domains, we have the explicit constant $\alpha(2, Q) = 1/Q$.*

For the second regularity theorem we need the definition of the singular set of f.

Definition 3.6 (Regular and singular points). A Dir-minimizing f is regular at a point $x \in \Omega$ if there exists a neighborhood B of x and Q analytic functions $f_i : B \to \mathbb{R}^n$ such that

$$f(y) = \sum_i \llbracket f_i(y) \rrbracket \qquad \text{for almost every } y \in B \tag{3.2}$$

and either $f_i(x) \neq f_j(x)$ for every $x \in B$, or $f_i \equiv f_j$. The singular set Σ_f of f is the complement of the set of regular points.

Theorem 3.7 (Estimate of the singular set). *Let f be* Dir*-minimizing. Then, the singular set Σ_f of f is relatively closed in Ω. Moreover, if $m = 2$, then Σ_f is at most countable, and if $m \geq 3$, then the Hausdorff dimension of Σ_f is at most $m - 2$.*

Note in particular the striking similarity between the estimate of the size of the singular set in the case of multiple valued Dir-minimizers and in that of area-minimizing currents. It will be discussed later that, even in the case of Dir-minimizers, there are singular solutions (which are no better than Hölder continuous).

Complete and self-contained proofs of these theorems can be found in [13]. The key idea beyond the estimate for the singular set is the celebrated frequency function (cp. with [13, Section 3.4]), which has been indeed used in a variety of different contexts in the theory of unique continuation of partial differential equations (see for instance the papers [20], [21]). This is the central tool of our proofs as well. However, our arguments manage much more efficiently the technical intricacies of the problem and some aspects of the theory are developed in further details. For instance, we present in [13, Section 3.1] the Euler-Lagrange conditions derived from first variations in a rather general form. This is to our knowledge the first time that these conditions appear somewhere in this generality.

Largely following ideas of [7] and of White, we improve the second regularity theorem to the following optimal statement for planar maps.

Theorem 3.8 (Improved estimate of the singular set). *Let f be* Dir*-minimizing and $m = 2$. Then, the singular set Σ of f consists of isolated points.*

This result was announced in [7]. However, to our knowledge the proof has never appeared so far. For a discussion of the optimality of these regularity results, we refer the reader to Section 5 below.

4. Metric Analysis

4.1. An intrinsic approach. One of the less satisfactory points of Almgren's theory is the heavy use of the Lipschitz maps $\boldsymbol{\xi}$ and $\boldsymbol{\rho}$. First of all, this makes the arguments often counterintuitive. Second, there is the obvious disturbing fact that, while several choices of $\boldsymbol{\xi}$ and $\boldsymbol{\rho}$ are possible, the objects of the study and the ultimate conclusions of the theory are totally independent of this choice. This fact has been pointed out for the first time in [24]. In the papers [24] and [25] the author made some progress in the program of making Almgren's theory "intrinsic", i.e. independent of the euclidean embedding.

As far as the theory of Dir-minimizers is concerned, this program has been completed in our paper [13]. This work also makes a clear link between Almgren's theory and the vast existing literature about metric analysis, metric geometry and general harmonic maps, which started with the pioneering papers [22], [30] and [5] (we refer the interested reader to [13, Section 4.1]).

The metric approach has several features:

- One first advantage is that it allows to separate "hard" and "soft" parts in Almgren's theory. Several conclusions can indeed be reached in a straightforward way by "abstract nonsense". Only few key points need deeply the structure of $\mathcal{A}_Q(\mathbb{R}^n)$ and some "hard" computations. By quickly discarding the minor points, the metric theory is a powerful tool to recognize plausible statements and crucial issues.

- A second advantage is the natural link to the metric theory of currents developed by Ambrosio and Kirchheim in [6]. This theory recovers many of the central theorems of Federer and Fleming's work [18] in a clean way and offers some new powerful tools (like the Jerrard-Soner BV estimates for the slicing theory). The reason why this connection is useful will be explored in detail in Section 6.

4.2. Intrinsic definition of the Dirichlet energy. The metric point of view relies upon the following alternative definitions of Dirichlet energy and Sobolev functions (cp. with the general theory developed in [5] and [34]; the careful reader will notice, however, that there is a crucial difference between the definition of Dirichlet energy in [34] and the one given below).

Definition 4.1 (Sobolev Q-valued functions). A measurable $f : \Omega \to \mathcal{A}_Q$ is in the Sobolev class $W^{1,p}$ $(1 \leq p \leq \infty)$ if there exist m functions $\varphi_j \in L^p(\Omega; \mathbb{R}^+)$ such that

(i) $x \mapsto \mathcal{G}(f(x), T) \in W^{1,p}(\Omega)$ for all $T \in \mathcal{A}_Q$;

(ii) $|\partial_j \mathcal{G}(f, T)| \leq \varphi_j$ a.e. in Ω for all $T \in \mathcal{A}_Q$ and for all $j \in \{1, \ldots, m\}$.

It is not difficult to show the existence of minimal functions $\tilde{\varphi}_j$ fulfilling (ii), i.e. such that, for any other φ_j satisfying (ii), $\tilde{\varphi}_j \leq \varphi_j$ a.e. (cp. with [13, Proposition 4.2]). Such "minimal bounds" will be denoted by $|\partial_j f|$ and we note that they are characterized by the following property (see again [13, Proposition 4.2]): for every countable dense subset $\{T_i\}_{i \in \mathbb{N}}$ of $\mathcal{A}_Q$ and for every $j = 1, \ldots, m$,

$$|\partial_j f| \;=\; \sup_{i \in \mathbb{N}} |\partial_j \mathcal{G}(f, T_i)| \quad \text{almost everywhere in } \Omega. \tag{4.1}$$

We are now ready to define the Dirichlet energy.

Definition 4.2. The function $|Df|^2$ is defined to be the sum of $|\partial_j f|^2$. The Dirichlet energy of $f \in W^{1,2}(U; \mathcal{A}_Q)$ is then defined by $\mathrm{Dir}(f, U) := \int_U |Df|^2$.

As already mentioned, this definition is equivalent to the one proposed in the previous section.

The paper [13] gives therefore two different approaches to the theorems stated in the previous section. One can follow a (considerably simpler) version of Almgren's "extrinsic" approach, exploiting the maps $\boldsymbol{\xi}$ and $\boldsymbol{\rho}$. Or one can use the intrinsic approach starting from the definitions above, without using the maps $\boldsymbol{\xi}$ and $\boldsymbol{\rho}$. However, proceeding further in Almgren's program for the regularity of area-minimizing currents, there is a point at which we have not been able to avoid these extrinsic maps (see Sections 6.4 and 8).

5. Higher Integrability of Dir Minimizers and Other Results

5.1. Multiple valued functions beyond Almgren. Many results of Almgren have been extended in several directions. In particular

- The papers [11], [25], [52], [53] extend some of Almgren's results to ambient spaces which are more general than the euclidean one;
- The papers [51], [54], [26] and [24] consider some other objects in the multiple valued setting (such as differential inclusions, geometric flows and quasiminima);
- The papers [31] and [12] extend some of Almgren's theorems to more general energy functionals.

5.2. Higher integrability. In this section we focus on a recent new contribution to the theory, which plays an important role in our derivation of the second main step in Almgren's program. Dir-minimizing functions enjoy higher integrability of the gradient. We believe that several intricate arguments and complicated constructions in Almgren's third chapter can be reinterpreted as rather particular cases of this key observation (see for instance [4, Section 3.20]). Surprisingly, this higher integrability can be proved in a very simple way by deriving a suitable reverse Hölder inequality and using a (nowadays) very standard version of the classical Gehring's Lemma.

Theorem 5.1 (Higher integrability of Dir-minimizers). *Let $\Omega' \subset\subset \Omega \subset\subset \mathbb{R}^m$ be open domains. Then, there exist $p > 2$ and $C > 0$ such that*

$$\|Du\|_{L^p(\Omega')} \leq C \, \|Du\|_{L^2(\Omega)} \quad \text{for every Dir-minimizing } u \in W^{1,2}(\Omega, \mathcal{A}_Q(\mathbb{R}^n)). \tag{5.1}$$

This theorem has been stated and proved for the first time in [14]. The relevant reverse Hölder inequality has been derived using a comparison argument and hence relying heavily on the minimality of the Dir-minimizers. A second

proof, exploiting the Euler-Lagrange conditions to give a Caccioppoli-type inequality, has been given in [44]. This last proof still uses the regularity theory for Dir-minimizers. However, this occurs only at one step: one could hope to remove this restriction and generalize the higher integrability to "critical" points of the Dirichlet energy (cp. with Section 8).

5.3. Optimality. In [44] a yet different proof for the planar case is proposed, yielding the optimal range of exponents p for which (5.1) holds. The optimality of this result, as well as the optimality of Theorems 3.5 and 3.8, is shown by another remarkable observation of Almgren. Besides giving area-minimizing currents, holomorphic varieties are locally graphs of Dir minimizing Q-valued functions. In [4, Section 2.20] Almgren proves this statement appealing to his powerful approximation results for area-minimizing currents (see Section 6 below). However this is unnecessary and a rather elementary proof can be found in [44].

6. Approximation of Area-minimizing Currents

After developing the theory of multiple valued functions, Almgren devotes the third chapter of his monograph to a suitable approximation theorem for area-minimizing currents, which is the multiple valued counterpart of the classical approximation theorem of De Giorgi in his proof of the excess-decay property.

6.1. Almgren's main approximation theorem. We start by giving the exact statement of Almgren's approximation result in the euclidean setting. Compared to the rest of the note, this part is rather technical. On the other hand, in order to get an understanding of Almgren's approximation theorem, a certain familiarity with the theory of currents can hardly be avoided.

Consider integer rectifiable m-dimensional currents T supported in some open cylinder $\mathcal{C}_r(y) = B_r(y) \times \mathbb{R}^n \subset \mathbb{R}^m \times \mathbb{R}^n$ and satisfying the following assumption:

$$\pi_\# T = Q \, [\![B_r(y)]\!] \qquad \text{and} \qquad \partial T = 0, \tag{6.1}$$

where $\pi : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ is the orthogonal projection and m, n, Q are fixed positive integers. In an informal language, the hypothesis (6.1) means that the current "covers" Q times the base of the cylinder.

We denote by $\mathfrak{e}_T$ the non-negative *excess measure* and by $\mathrm{Ex}(T, \mathcal{C}_r(y))$ the *cylindrical excess*, respectively defined by

$$\mathfrak{e}_T(A) := \mathbf{M}\big(T \llcorner (A \times \mathbb{R}^n)\big) - Q\,|A| \quad \text{for every Borel } A \subset B_r(y), \tag{6.2}$$

$$\mathrm{Ex}(T, \mathcal{C}_r(y)) := \frac{\mathfrak{e}_T(B_r(x))}{|B_r(x)|} = \frac{\mathfrak{e}_T(B_r(x))}{\omega_m r^m}. \tag{6.3}$$

Though it is not apparent from the definition given above, the cylindrical excess bears some similarities with the spherical excess.

Theorem 6.1. *There exist constants $C, \delta, \varepsilon_0 > 0$ with the following property. Let T be an area-minimizing, integer rectifiable m-dimensional current in the cylinder $\mathcal{C}_4$ which satisfies* (6.1)*. If $E = \mathrm{Ex}(T, \mathcal{C}_4) < \varepsilon_0$, then there exist a Q-valued function $f \in \mathrm{Lip}(B_1, \mathcal{A}_Q(\mathbb{R}^n))$ and a closed set $K \subset B_1$ such that*

$$\mathrm{Lip}(f) \leq C E^{\delta}, \tag{6.4}$$

$$\mathrm{graph}(f|_K) = T \llcorner (K \times \mathbb{R}^n) \quad \textit{and} \quad |B_1 \setminus K| \leq C E^{1+\delta}, \tag{6.5}$$

$$\left| \mathbf{M}\big(T \llcorner \mathcal{C}_1\big) - Q\,\omega_m - \int_{B_1} \frac{|Df|^2}{2} \right| \leq C\, E^{1+\delta}. \tag{6.6}$$

An interesting aspect which makes the proof of Theorem 6.1 quite hard is the gain of a small power E^{δ} in the three estimates (6.4), (6.5) and (6.6). Observe that the usual approximation theorems stated commonly in the literature, which cover the case $Q = 1$ and "stationary currents" (in fact, stationary integral varifolds), are stated with $\delta = 0$. On the other hand, the gain of Theorem 6.1 plays a crucial role in some of the estimates needed for the third main step of Almgren's program, i.e. the "construction of the center manifold" (cp. with Section 7).

6.2. Higher integrability for area-minimizing currents.

The note [14] provides a different, much simpler proof of Almgren's theorem. A key point is a higher integrability estimate for the Lebesgue density $\mathfrak{d}_T$ of the measure $\mathfrak{e}_T$, called the *excess density*,

$$\mathfrak{d}_T(x) := \limsup_{s \to 0} \frac{\mathfrak{e}_T(B_s(x))}{\omega_m\, s^m}.$$

Theorem 6.2. *There exist constants $p > 1$ and $C, \varepsilon > 0$ with the following property. Assume T is an area-minimizing, integer rectifiable current of dimension m. If T satisfies* (6.1) *and $E = \mathrm{Ex}(T, \mathcal{C}_4) < \varepsilon$, then*

$$\int_{\{\mathfrak{d} \leq 1\} \cap B_2} \mathfrak{d}^p \leq C\, E^p. \tag{6.7}$$

This estimate, which can be thought as the "current counterpart" of Theorem 5.1, is not explicitly stated in [4], but it can be deduced from some of the arguments therein. These arguments, which include quite elaborate constructions and use several intricate covering algorithms, are the most involved part of Almgren's proof.

One comment is in order. In the case $Q = 1$ we know a posteriori that T coincides with the graph of a $C^{1,\alpha}$ function over B_2 (cp. with Theorem 2.1). However, the branching phenomenon makes Theorem 6.2 much more interesting in the higher codimension, since essentially it cannot be improved (except in the sense of optimizing the exponent p and the constant C). Consider in particular the following example. Let η be a rather small constant and T be the current

associated to the holomorphic variety $\{z^2 = \eta w\} \subset \mathbb{C}^2 = \mathbb{R}^4$. Set $\mathcal{C}_4 := \{|w| < 4\}$ and $Q = 2$. If η is chosen very small compared to ε, then T satisfies all the assumptions of Theorem 6.1. On the other hand, the corresponding function $\mathfrak{d}_T$ does not belong to L^2 and one can easily check that estimate (6.7) does not hold if $p \geq 2$.

6.3. Some new techniques coming from metric analysis. The main contribution of [14] is to give a much shorter and conceptually clearer derivation of (6.7) (in fact, since Theorem 6.2 is not stated by Almgren, the real point is to establish Theorem 6.6 below, which however is trivially equivalent). Moreover, in [14] we introduce several new ideas. In particular:

(i) we introduce a powerful maximal function truncation technique to approximate general integer rectifiable currents with multiple valued functions;

(ii) we give a simple compactness argument to conclude directly a first harmonic approximation of T;

(iii) we give a new proof of the existence of Almgren's "almost projections" $\boldsymbol{\rho}^\star$.

In the rest of this section we look more closely at these ideas.

Given a normal m-current T, following [6] we can view the slice map $x \mapsto \langle T, \pi, x\rangle$ as a BV function taking values in the space of 0-dimensional currents (endowed with the flat metric). Indeed, by a key estimate of Jerrard and Soner (see [6] and [29]), the total variation of the slice map is controlled by the mass of T and ∂T. In the same vein, following [13], Q-valued functions can be viewed as Sobolev maps into the space of 0-dimensional currents. These two points of view can be combined with standard maximal function truncation arguments to develop a powerful and simple Lipschitz approximation technique, which gives a systematic tool to find graphical approximations of integer rectifiable currents.

To give a more precise idea of this method, we introduce the maximal function of the excess measure of a current T (satisfying (6.1)):

$$M_T(x) := \sup_{B_s(x)\subset B_r(y)} \frac{\mathfrak{e}_T(B_s(x))}{\omega_m\, s^m} = \sup_{B_s(x)\subset B_r(y)} \mathrm{Ex}(T, \mathcal{C}_s(x)).$$

Our main approximation result is the following and relies on an improvement of the usual Jerrard–Soner estimate.

Proposition 6.3 (Lipschitz approximation). *There exist constants $c, C > 0$ with the following property. Let T be an integer rectifiable m-current in $\mathcal{C}_{4s}(x)$ satisfying* (6.1) *and let $\eta \in (0, c)$ be given. Set $K := \big\{M_T < \eta\big\} \cap B_{3s}(x)$. Then,*

there exists $u \in \mathrm{Lip}(B_{3s}(x), \mathcal{A}_Q(\mathbb{R}^n))$ *such that* $\mathrm{graph}(u|_K) = T \llcorner (K \times \mathbb{R}^n)$, $\mathrm{Lip}(u) \leq C\,\eta^{\frac{1}{2}}$ *and*

$$|B_{3s}(x) \setminus K| \leq \frac{C}{\eta}\,\mathbf{e}_T\big(\{M_T > \eta/2\}\big). \tag{6.8}$$

In the rest of this section, we will often choose $\eta = E^{2\alpha}$ $(= \mathrm{Ex}(T, \mathcal{C}_{4s}(x))^{2\alpha})$, for some $\alpha \in (0, (2m)^{-1})$. The map u given by Proposition 6.3 will then be called *the* E^α*-Lipschitz* (or briefly *the Lipschitz*) approximation of T in $\mathcal{C}_{3s}(x)$. We therefore conclude the following estimates:

$$\mathrm{Lip}(u) \leq C\,E^\alpha, \tag{6.9}$$

$$|B_{3s}(x) \setminus K| \leq C\,E^{-2\,\alpha}\,\mathbf{e}_T\big(\{M_T > E^{2\,\alpha}/2\}\big), \tag{6.10}$$

$$\int_{B_{3s}(x)\setminus K} |Du|^2 \leq \mathbf{e}_T\big(\{M_T > E^{2\,\alpha}/2\}\big). \tag{6.11}$$

In particular, the function f in Theorem 6.1 is given by the E^α-Lipschitz approximation of T in $\mathcal{C}_1$, for a suitable choice of α.

The second step in the proof of Theorem 6.2 is a compactness argument which shows that, when T is area-minimizing, the approximation f is close to a Dir-minimizing function w, with an $o(E)$ error.

Theorem 6.4 ($o(E)$-improvement). *Let* $\alpha \in (0, (2m)^{-1})$. *For every* $\eta > 0$, *there exists* $\varepsilon_1 = \varepsilon_1(\eta) > 0$ *with the following property. Let* T *be a rectifiable, area-minimizing* m*-current in* $\mathcal{C}_{4s}(x)$ *satisfying* (6.1). *If* $E \leq \varepsilon_1$ *and* f *is the* E^α*-Lipschitz approximation of* T *in* $\mathcal{C}_{3s}(x)$, *then*

$$\int_{B_{2s}(x)\setminus K} |Df|^2 \leq \eta\,\mathbf{e}_T(B_{4s}(x)), \tag{6.12}$$

and there exists a Dir*-minimizing* $w \in W^{1,2}(B_{2s}(x), \mathcal{A}_Q(\mathbb{R}^n))$ *such that*

$$\int_{B_{2s}(x)} \mathcal{G}(f,w)^2 + \int_{B_{2s}(x)} \big(|Df| - |Dw|\big)^2 \leq \eta\,\mathbf{e}_T(B_{4s}(x)). \tag{6.13}$$

This theorem is the multi-valued analog of De Giorgi's harmonic approximation, which is ultimately the heart of all the regularity theories for minimal surfaces. Our compactness argument is, to our knowledge, new (even for $n = 1$) and particularly robust. Indeed, we expect it to be useful in more general situations.

Next, Theorems 6.4 and 5.1 imply the following key estimate, which leads to Theorem 6.2 via an elementary "covering and stopping radius" argument.

Proposition 6.5. *For every* $\kappa > 0$, *there is* $\varepsilon_2 > 0$ *with the following property. Let* T *be an integer rectifiable, area-minimizing current in* $\mathcal{C}_{4s}(x)$ *satisfying* (6.1). *If* $E \leq \varepsilon_2$, *then*

$$\mathbf{e}_T(A) \leq \kappa\,E s^m \quad \textit{for every Borel } A \subset B_s(x) \textit{ with } |A| \leq \varepsilon_2 |B_s(x)|. \tag{6.14}$$

Using now Theorem 6.2, we can prove the most important estimate contained in Chapter 3 of [4].

Theorem 6.6. *There exist constants $\sigma, C > 0$ with the following property. Let T be an area-minimizing, integer rectifiable T of dimension m in $\mathcal{C}_4$. If T satisfies* (6.1) *and $E = \mathrm{Ex}(T, \mathcal{C}_4) < \varepsilon_0$, then*

$$\mathfrak{e}_T(A) \leq C\, E\left(E^\sigma + |A|^\sigma\right) \quad \textit{for every Borel } A \subset B_{4/3}. \tag{6.15}$$

6.4. Almgren's "almost projection" $\boldsymbol{\rho^\star}$. The proof of Theorem 6.6 is then the only part where we follow essentially Almgren's strategy. The main point is to estimate the size of the set over which the graph of the Lipschitz approximation f differs from T. As in many standard references, in the case $Q = 1$ this is achieved comparing the mass of T with the mass of the graph of $f * \rho_{E^\omega}$, where ρ is a smooth convolution kernel and $\omega > 0$ a suitably chosen constant (this idea is, essentially, already contained in De Giorgi's original proof).

However, for $Q > 1$, the space $\mathcal{A}_Q(\mathbb{R}^n)$ is not linear and we cannot regularize f by convolution. To bypass this problem, we follow Almgren and view $\mathcal{A}_Q$ as a subset of a large Euclidean space (via the biLipschitz embedding $\boldsymbol{\xi}$). We can then take the convolution of the map $\boldsymbol{\xi} \circ f$ and project it back on the set $\boldsymbol{\xi}(\mathcal{A}_Q)$. However, in order to do this efficiently in terms of the energy, we need an "almost" projection, denoted by $\boldsymbol{\rho}^\star_\mu$, which is almost 1-Lipschitz in the μ-neighborhood of $\boldsymbol{\xi}(\mathcal{A}_Q(\mathbb{R}^n))$ (μ is a parameter which must be tuned accordingly). At this point Theorem 6.2 enters in a crucial way in estimating the size of the set where the regularization of $\boldsymbol{\xi} \circ f$ is far from $\boldsymbol{\xi}(\mathcal{A}_Q(\mathbb{R}^n))$.

The maps $\boldsymbol{\rho}^\star_\mu$ are slightly different from Almgren's almost projections, but similar in spirit. In [14] we propose on original argument for the construction of $\boldsymbol{\rho}^\star_\mu$. One advantage of this argument is that it yields more explicit estimates in terms of the crucial parameter μ. As mentioned earlier, this is so far the only stage where we cannot avoid Almgren's extrinsic maps. It would be of interest to develop a more intrinsic approximation procedure, bypassing this "convolution and projection" technique (cp. Section 8 below).

7. Center Manifold: A Case Study

The fourth chapter of the big regularity paper (and roughly half of this monograph) is devoted to the construction of the so called "center manifold". In that chapter Almgren succeeds in constructing a $C^{3,\alpha}$ regular surface, which he calls center manifold and, roughly speaking, approximates the "average of the sheets of the current" (we refer to [4] for further details) in a neighborhood of a branching point. In the model example of Remark 2.2, the "ideal center manifold" would be the plane $\{z = 0\}$.

Essentially, the center manifold plays the same role of the barycenters of the measures $u(x)$ when u is a Q-valued map. In the latter example, it is rather straightforward to prove that the resulting "average function" is a classical

harmonic function (see for example [13, Lemma 3.23]). Unfortunately for the case of area-minimizing current, due to the "nonlinear nature" of the problem, there is no obvious PDE allowing for a similar conclusion.

7.1. Higher regularity "without PDEs". In the introduction of [4] Almgren observes that, in the case $Q = 1$, the center manifold coincides necessarily with the current itself, thus implying directly its $C^{3,\alpha}$ regularity. Compared to the usual proofs, this is rather striking. In fact, after proving Theorem 2.1, the "usual" regularity theory proceeds further by deriving the well-known Euler–Lagrange equations for the function f. It then turns out that f solves a system of elliptic partial differential equations and the Schauder theory implies that f is smooth (in fact analytic, using the classical result by Hopf [28]).

The corollary of Almgren's construction is that the $C^{3,\alpha}$ regularity can be concluded without appealing to "nonparametric techniques". In the note [15] we give a simple direct proof of this remark, essentially following Almgren's strategy for the construction of the center manifold in the case $Q = 1$. Though in a very simplified situation, this model case retains several key estimates of Almgren's construction. For instance it makes transparent the fundamental role played by the E^{δ}-gain in the estimates of the Approximation Theorem 6.1.

Our hope is that this will be a first step in the full understanding of Almgren's result. It is worthwhile to notice that, compared to the extremely long construction of the center manifold, the last portion of [4], containing the concluding arguments of Almgren's regularity theorem for area-minimizing integral currents, is much shorter. The construction of the center manifold seems the last big obstacle which needs to be overcome in order to understand the full regularity results of Almgren and Chang.

It is of a certain interest to notice that this "higher regularity" result stops a little after three derivatives. It does not seem possible, for instance, to get an estimate for the C^4 norm. In the proof presented in [15], this is quite transparent. In some sense, one can think of Almgren's strategy as an extremely careful approximation of the current obtained by pasting together (suitably rotated) graphs of harmonic functions.

One reason for the $C^{3,\alpha}$ estimate might be the fact that the Dirichlet energy is a quite accurate approximation of the area functional. Loosely speaking, one can think of De Giorgi's theorem as a consequence of the fact that the harmonic functions are first order expansions of solutions to the minimal surfaces. One gains almost 2 derivatives in this way (a careful look at the proof of Theorem 2.1 would show that it works for every $\beta < 1$, cp. with the Appendix of [15]). Taking the Taylor expansion to the next level, it turns out that harmonic functions approximate solutions of the minimal surface equations even "to the next order". To illustrate this phenomenon, consider the simpler situation of a surface of codimension 1, given by the graph of a Lipschitz function f. The key ingredient in De Giorgi's argument for the excess-decay is the following

observation on the integrand of the area functional:

$$I(\nabla f) \;:=\; \sqrt{1+|\nabla f|^2} \;=\; 1+\frac{1}{2}|\nabla f|^2+o(|\nabla f|^2)\,.$$

However, the Taylor expansion yields a much more precise information:

$$I(\nabla f) \;=\; 1+\frac{1}{2}|\nabla f|^2+O(|\nabla f|^4)\,. \tag{7.1}$$

The identity (7.1) is correct also in higher codimension.

8. Open Problems

In this section we collect a list of open problems on Q-valued functions. As already mentioned, there are several directions in which Almgren's theory could be extended, in particular in generalizing it to non-euclidean ambient spaces. However, in this list we have decided to focus on the euclidean setting and on problems which would deliver new information rather than generalizing existing theorems to different contexts. Many of these problems have been proposed by Almgren and the reader might find them in the collection [1].

(1) In the proof of Theorem 3.7 a pivotal role is played by the so called "tangent functions". The key idea (which ultimately might be regarded as the most important discovery of Almgren) is that, when suitably rescaling a Dir-minimizer in a neighborhood of a singularity, the resulting maps converge, up to subsequences, to Dir-minimizers which are radially homogeneous. This theorem is achieved through the monotonicity of the celebrated frequency function, which in this context plays the same role of the monotonicity formula for area-minimizing currents.

The uniqueness of the "blow-up" at a singularity is not known, except for the planar maps (see [13, Theorem 5.3], where it is proved before Theorem 3.8 exploiting some ideas of [7]; assuming Theorem 3.8, this uniqueness is an obvious consequence of the considerations in [32]). Almgren suggests that a relevant role in this problem might be played by the techniques developed in [40] (cp. with [1, Problem 5.6]).

(2) A tentative conjecture is that the singular set of a Dir-minimizer map on an m–dimensional domain should have (locally) finite $\mathcal{H}^{m-2}$ measure and be rectifiable. This is only known to hold in the case $m = 2$ (cp. with Theorem 3.8).

(3) In [1, Problem 5.5] Almgren asks whether the graph of a Dir-minimizer is always a real analytic set. To our knowledge this is unknown even in the case of planar maps, where rather detailed information is available (after combining Theorem 3.8 with the results of [32]).

(4) It would be interesting to get other examples of Dir-minimizers. To our knowledge, no other systematic class of examples is known apart from

that of holomorphic varieties (cp. with the discussion in Section 5.3). Is there any other similar class that one could derive from other calibrated geometries?

(5) Essentially nothing is known if we replace the minimizing property with stationarity. Different notions of stationary maps are possible, due to the difference between inner and outer variations (cp. with [13, Section 3.1]) and to the possibility of introducing more general type of deformations. Does the singular set have measure zero? It is easy to see that there are maps which are stationary with respect to both inner and outer variations and have a singular set of dimension $m-1$. Does the singular set have dimension $m-2$ if the map is stationary with respect to any one-parameter family of deformations (cp. with [4, Problem 5.5])?

(6) Very little is known if we change the Dirichlet energy. The paper [12] shows the existence of a large class of semicontinuous functionals. If we restrict to planar maps and quadratic (semicontinuous) functionals, the only information available for minimizers is the Hölder continuity (proved in [31]).

(7) Are Dir-minimizers continuous, or ever Hölder, up to the boundary, if the boundary data are sufficiently regular? The only known result is the continuity for 2–dimensional domains (proved in [50]).

(8) Can one avoid the map $\boldsymbol{\rho}^\star_\mu$ in the proof of Theorem 6.1? Another way to phrase this question is the following. Is there an "intrinsic" efficient smoothing procedure for Q-valued functions? So far the following are the only two available techniques:

- The (intrinsic) maximal function truncation argument which allows to approximate general Q-valued functions in $W^{1,p}$ with Lipschitz maps.
- Almgren's extrinsic smoothing: the given map u is transformed into a Euclidean map $\boldsymbol{\xi} \circ u$; this map is than regularized (for instance with a convolution) and, to produce again a Q-valued map, the regularization is projected on the set $\boldsymbol{\xi}(\mathcal{A}_Q)$.

The latter yields efficient estimates when dealing with the Dirichlet energies of the corresponding maps. We do not know of any intrinsic method to achieve regularizations with the same estimates.

References

[1] *Some open problems in geometric measure theory and its applications suggested by participants of the 1984 AMS summer institute.* Edited by J. E. Brothers. Proc. Sympos. Pure Math. **44**, Amer. Math. Soc., Providence, RI, USA, 1986.

[2] W.K. Allard, *On the first variation of a varifold*, Ann. of Math. (2) **95** (1972), 417–491.

[3] F.J. Almgren, Jr., *Some interior regularity theorems for minimal surfaces and an extension of Bernstein's theorem*, Ann. of Math. (2) **84** (1966), 277–292.

[4] F.J. Almgren, Jr., *Almgren's big regularity paper*, World Scientific, River Edge, NJ, USA, 2000.

[5] L. Ambrosio, *Metric space valued functions of bounded variation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **17** (1990), 439–478.

[6] L. Ambrosio, B. Kirchheim, *Currents in metric spaces*, Acta Math. **185** (2000), 1–80.

[7] S.X. Chang, *Two-dimensional area-minimizing currents are classical minimal surfaces*, J. Amer. Math. Soc. **1** (1988), 699–788.

[8] E. Bombieri, E. De Giorgi, E. Giusti, *Minimal cones and the Bernstein problem*, Invent. Math. **7** (1969), 243–268.

[9] E. De Giorgi, *Frontiere orientate di misura minima*, Editrice Tecnico Scientifica, Pisa, Italy, 1961.

[10] E. De Giorgi, *Una estensione del teorema di Bernstein*, Ann. Scuola Norm. Sup. Pisa, Ser 3 **19** (1965), 79–85.

[11] C. De Lellis, R. Grisanti, P. Tilli, *Regular selection for multiple valued functions*, Ann. Mat. Pura Appl. (4) **183** (2004), 79–95.

[12] C. De Lellis, M. Focardi, E. Spadaro, *Lower semicontinuous functionals on Almgren's multiple valued functions*, Preprint (2009).

[13] C. De Lellis, E. Spadaro, *Almgren's Q-valued functions revisited*, to appear in Mem. AMS.

[14] C. De Lellis, E. Spadaro, *Higher integrability and approximation of minimal currents*, Preprint (2009).

[15] C. De Lellis, E. Spadaro, *Center manifold: a case study*, Preprint (2010).

[16] H. Federer, *Geometric measure theory*, Die Grundlehren der mathematischen Wissenschaften, Band 153, Springer-Verlag New York Inc., New York, 1969.

[17] H. Federer, *The singular set of area-minimizing rectifiable currents with codimension* 1 *and of area-minimizing flat chains modulo two with arbitrary codimension*, Bull. AMS **76** (1970), 767–771.

[18] H. Federer, W. Fleming, *Normal and integral currents*, Ann. of Math. (2) **72** (1960), 458–520.

[19] W. Fleming, *On the oriented Plateau problem*, Rend. Circ. Mat. Palermo (2) **11** (1962), 69–90.

[20] N. Garofalo, F.H. Lin, *Monotonicity properties of variational integrals, A_p weights and unique continuation*, Indiana Univ. Math. J. **35** (1986), 245–268.

[21] N. Garofalo, F.H. Lin, *Unique continuation for elliptic operators: a geometric-variational approach*, Comm. Pure Appl. Math. **40** (1987), 347–366.

[22] M. Gromov, R. Schoen, *Harmonic maps into singular spaces and p-adic superrigidity for lattices in groups of rank one*, Inst. Hautes tudes Sci. Publ. Math. **76** (1992), 165–246.

[23] E. Giusti, *Minimal surfaces and functions of bounded variation*, Birkhäuser Verlag, Basel, 1984.

[24] J. Goblet, *A selection theory for multiple-valued functions in the sense of Almgren*, Ann. Acad. Sci. Fenn. Math. **31** (2006), 347–366.

[25] J. Goblet, *Lipschitz extension of multiple valued Banach-valued functions in the sense of Almgren*, Houston J. Math. **35** (2009), 223–231.

[26] J. Goblet, W. Zhu, *Regularity of Dirichlet nearly minimizing multiple-valued functions*, J. Geom. Anal. **18** (2008), 765–794.

[27] R. Harvey, H.B. Lawson, Jr., *Calibrated geometries*, Acta Math. **148** (1982), 47–157.

[28] E. Hopf, *Über den funktionalen, insbesondere den analytischen Charakter der Lösungen ellptischer Differentialgleichungen zweiter Ordnung*, Math. Z. **34** (1932), 194–233.

[29] R. Jerrard, H.M. Soner, *Functions of bounded higher variation*, Indiana Univ. Math. J. **51** (2002), 645–677.

[30] N. Korevaar, R. Schoen, *Sobolev spaces and harmonic maps for metric space targets*, Comm. Anal. Geom. **1** (1993), 561–659.

[31] P. Mattila, *Lower semicontinuity, existence and regularity theorems for elliptic variational integrals of multiple valued functions*, Trans. Amer. Math. Soc. **280** (1983), 589–610.

[32] M. Micallef, B. White, *The structure of branch points in minimal surfaces and in pseudoholomorphic curves*, Ann. of Math. (2) **141** (1995), 35–85.

[33] F. Morgan, *On the singular structure of two-dimensional area minimizing surfaces in* $\mathbb{R}^n$, Math. Ann. **261** (1982), 101–110.

[34] Y.G. Reshetnyak, *Sobolev classes of functions with values in a metric space*, Sibirsk. Math. Zh. **38** (1997), 657–675.

[35] T. Rivière, *A lower-epiperimetric inequality for area-minimizing surfaces*, Comm. Pure Appl. Math. **57** (2004), 1673–1685.

[36] T. Rivière, G. Tian, *The singular set of J-holomorphic maps into projective algebraic varieties*, J. Reine Angew. Math. **570** (2004), 47–87.

[37] T. Rivière, G. Tian, *The singular set of* $1-1$ *integral currents*, Ann. of Math. (2) **169** (2009), 741–794.

[38] R. Schoen, L. Simon, *Regularity of stable minimal hypersurfaces*, Comm. Pure Appl. Math. **34** (1981), 741–797.

[39] L. Simon, *Lectures on geometric measure theory*, Proceedings of the Centre for mathematical analysis, Australian National University, Canberra, 1983.

[40] L. Simon, *Asymptotics for a class of nonlinear evolution equations, with applications to geometric problems*, Ann. of Math. (2) **118** (1983), 525–571.

[41] L. Simon, *Rectifiability of the singular sets of multiplicity* 1 *minimal surfaces and energy minimizing maps*, Surveys in differential geometry **2** (1995), 246–305.

[42] L. Simon, *Theorems on the regularity and singularity of minimal surfaces and harmonic maps*, Lectures on geometric variational problems (Sendai, 1993), 115–150, Springer, Tokyo, 1996.

[43] J. Simons, *Minimal varieties in Riemannian manifolds*, Ann. of Math. (2) **88** (1968), 62–105.

[44] E.N. Spadaro, *Complex varieties and higher integrability of* Dir-*minimizing* Q*-valued functions*, to appear in Manuscripta Mathematica.

[45] C.H. Taubes, *Seiberg Witten and Gromov invariants for symplectic* 4*-manifolds*, First International Press Lecture Series, 2. International Press, Somerville, MA, USA, 2000.

[46] C. Villani, *Topics in optimal transportation*, Graduate studies in mathematics, vol. 58, AMS, Providence, RI, USA, 2003.

[47] B. White, *Tangent cones to two-dimensional area-minimizing integral currents are unique*, Duke Math. J. **50** (1983), 143–160

[48] N. Wickramasekera, *A general regularity theory for stable codimension 1 integral varifolds*, Preprint (2009).

[49] W. Wirtinger, *Eine Determinantenidentität und ihre Anwendung auf analytische Gebilde und Hermitsche Maßbestimmung*, Monatsh. f. Math. u. Physik **44** (1936), 343–365.

[50] W. Zhu, *Two-dimensional multiple-valued Dirichlet minimizing functions*, Comm. Partial Differential Equations **33** (2008), 1847–1861.

[51] W. Zhu, *Analysis on the metric space* $\mathcal{Q}$, Preprint (2006).

[52] W. Zhu, *A regularity theory for multiple-valued Dirichlet minimizing maps*, Preprint (2006).

[53] W. Zhu, *A theorem on the frequency function for multiple-valued Dirichlet minimizing functions*, Preprint (2006).

[54] W. Zhu, *An energy reducing flow for multiple-valued functions*, Preprint (2006).

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

New Entire Solutions to Some Classical Semilinear Elliptic Problems

Manuel del Pino*

Abstract

This paper deals with the construction of solutions to autonomous semilinear elliptic equations considered in entire space. In the absence of space dependence or explicit geometries of the ambient space, the point is to unveil internal mechanisms of the equation that trigger the presence of families of solutions with interesting concentration patterns. We discuss the connection between minimal surface theory and entire solutions of the Allen-Cahn equation. In particular, for dimensions 9 or higher, we build an example that provides a negative answer to a celebrated question by De Giorgi for this problem. We will also discuss related results for the (actually more delicate) standing wave problem in nonlinear Schrödinger equations and for sign-changing solutions of the Yamabe equation.

Mathematics Subject Classification (2010). Primary 35J60; Secondary 35B25, 35B33.

Keywords. Allen-Cahn equation, standing waves for NLS, Yamabe equation.

1. Introduction

Understanding the entire solutions of nonlinear elliptic equations in $\mathbb{R}^N$ such as

$$\Delta u + f(u) = 0 \quad \text{in } \mathbb{R}^N, \tag{1.1}$$

is a basic problem in PDE research. This is the context of various classical results in literature like the Gidas-Ni-Nirenberg theorems on radial symmetry, Liouville type theorems, or the achievements around De Giorgi's conjecture. In those results, the geometry of level sets of the solutions turns out to be a

*This work has been supported by grants Fondecyt 1070389, Anillo ACT125 and Fondo Basal CMM.

Departamento de Ingeniería Matemática and CMM, Universidad de Chile, Casilla 170, Correo 3, Santiago, Chile. E-mail: delpino@dim.uchile.cl.

posteriori very simple (planes or spheres). On the other hand, problems of the form (1.1) with nonlinearities recurrent the literature, do have solutions with more interesting patterns, and the structure of their solution sets has remained mostly a mystery.

In many studies, problems like (1.1) are considered involving explicit dependence on the space variable, or on a manifold or in a domain in $\mathbb{R}^N$ under boundary conditions. Topological and geometric features of the domain are often characteristic that trigger the presence of interesting solutions, whose precise features can be analyzed when some singular perturbation parameter is involved. In the absence of space inhomogeneity or geometry of the ambient space, as in the "clean" equation (1.1), it is less clear which internal mechanisms of the equation are behind complex patterns in the solution set, whose richness may be nearly impossible to fully grasp.

In this paper we consider specific problems of the form (1.1) and describe recent results on existence of families of solutions, depending on parameters, that exhibit interesting asymptotic patterns linked to geometric objects in entire space. We consider the following three classical problems:

1. The Allen-Cahn equation,

$$\Delta u + u - u^3 = 0 \quad \text{in } \mathbb{R}^N .$$

2. The standing wave problem for the (focusing) nonlinear Schrödinger equation

$$\Delta u + |u|^{p-1}u - u = 0 \quad \text{in } \mathbb{R}^N .$$

3. The Yamabe equation

$$\Delta u + |u|^{\frac{4}{N-2}}u = 0 \quad \text{in } \mathbb{R}^N , \quad N \geq 3.$$

Sections 1 to 5, will be devoted to discuss the Allen Cahn equation. We will describe a link between entire minimal surfaces and solutions to the equation which have a nodal set close to large dilations near such a surface, while approaching ± 1 away from it, in particular answering negatively a long-standing question by De Giorgi in dimensions $N \geq 9$. We shall describe in Section 6 parallels and related results for the other two problems, which are in turn more delicate.

2. The Allen-Cahn Equation

The Allen-Cahn equation in $\mathbb{R}^N$ is the semilinear elliptic problem

$$\Delta u + u - u^3 = 0 \quad \text{in } \mathbb{R}^N . \tag{2.1}$$

Originally formulated in the description of bi-phase separation in fluids [14] and ordering in binary alloys [3], Equation (2.1) has received extensive mathematical

study. It is a prototype for the modeling of phase transition phenomena in a variety of contexts.

Introducing a small positive parameter ε and writing $v(x) := u(\varepsilon^{-1}x)$, we get the scaled version of (2.1),

$$\varepsilon^2 \Delta v + v - v^3 = 0 \quad \text{in } \mathbb{R}^N . \tag{2.2}$$

On every bounded domain $\Omega \subset \mathbb{R}^N$, (2.1) is the Euler-Lagrange equation for the action functional

$$J_\varepsilon(v) = \int_\Omega \frac{\varepsilon}{2} |\nabla v|^2 + \frac{1}{4\varepsilon}(1 - v^2)^2.$$

We observe that the constant functions $v = \pm 1$ minimize J_ε. They are idealized as two *stable phases* of a material in Ω. It is of interest to analyze configurations in which the two phases coexist. These states are represented by stationary points of J_ε, or solutions v_ε of Equation (2.2), that take values close to $+1$ in a subregion of Ω of and -1 in its complement. Modica and Mortola [64] and Modica [63], established that a family of local minimizers v_ε of J_ε for which

$$\sup_{\varepsilon > 0} J_\varepsilon(v_\varepsilon) < +\infty \tag{2.3}$$

must satisfy as $\varepsilon \to 0$, after passing to a subsequence,

$$v_\varepsilon \to \chi_\Lambda - \chi_{\Omega \setminus \Lambda} \quad \text{in } L^1_{loc}(\Omega). \tag{2.4}$$

Here Λ is an open subset of Ω with $\Gamma = \partial\Lambda \cap \Omega$ having *minimal perimeter*, being therefore a (generalized) minimal surface. Moreover,

$$J_\varepsilon(v_\varepsilon) \to \frac{2}{3}\sqrt{2}\mathcal{H}^{N-1}(\Gamma). \tag{2.5}$$

2.1. Formal asymptotic behavior of v_ε. Let us argue formally to obtain an idea on how a solution v_ε of Equation (2.2) with uniformly bounded energy (2.3) should look like near a limiting interface Γ. Let us assume that Γ is a smooth hypersurface and let ν designate a choice of its unit normal. Points δ-close to Γ can be uniquely represented as

$$x = y + z\nu(y), \quad y \in \Gamma, \ |z| < \delta \tag{2.6}$$

A well known formula for the Laplacian in these coordinates reads as follows.

$$\Delta_x = \partial_{zz} + \Delta_{\Gamma^z} - H_{\Gamma_z}\partial_z \tag{2.7}$$

Here

$$\Gamma^z := \{y + z\nu(y) \ / \ y \in \Gamma\}.$$

Δ_{Γ^z} is the Laplace-Beltrami operator on Γ^z acting on functions of the variable y, and H_{Γ^z} designates its mean curvature. Let $k_1, \ldots, k_N$ denote the principal curvatures of Γ. Then we have the validity of the expression

$$H_{\Gamma^z} = \sum_{i=1}^{N} \frac{k_i}{1 - zk_i}. \tag{2.8}$$

It is reasonable to assume that the solution has uniform smoothness in the y-direction, while in the transition direction z, elliptic estimates applied to the transformed equation (2.1) yield uniform smoothness in the variable $\zeta = \varepsilon^{-1}z$. The equation for $v_\varepsilon(y, \zeta)$ then reads

$$\varepsilon^2 \Delta_{\Gamma^{\varepsilon\zeta}} v_\varepsilon - \varepsilon H_{\Gamma^{\varepsilon\zeta}}(y)\, \partial_\zeta v_\varepsilon + \partial_\zeta^2 v_\varepsilon + v_\varepsilon - v_\varepsilon^3 = 0, \quad y \in \Gamma, \quad |\zeta| < \delta\varepsilon^{-1}. \tag{2.9}$$

We shall make two strong assumptions:

1. The zero-level set of v_ε lies within a $O(\varepsilon^2)$-neighborhood of Γ, that is on the region $|\zeta| = O(\varepsilon)$ and $\partial_\tau v_\varepsilon > 0$ on this nodal set, and

2. $v_\varepsilon(y, \zeta)$ can be expanded in powers of ε as

$$v_\varepsilon(y, \zeta) = v_0(y, \zeta) + \varepsilon v_1(y, \zeta) + \varepsilon^2 v_2(y, \zeta) + \cdots \tag{2.10}$$

for smooth coefficients bounded, with bounded derivatives. We observe also that

$$\int_\Gamma \int_{-\delta/\varepsilon}^{\delta/\varepsilon} [\, \frac{1}{2}|\partial_\zeta v_\varepsilon|^2 + \frac{1}{4}(1 - v_\varepsilon^2)^2 \,]\, d\zeta\, d\sigma(y) \le J_\varepsilon(v_\varepsilon) \le C \tag{2.11}$$

Substituting Expression (2.10) in Equation (2.9), using the first assumption, and letting $\varepsilon \to 0$, we get

$$\begin{aligned} \partial_\zeta^2 v_0 + v_0 - v_0^3 = 0, \quad & (y, \zeta) \in \Gamma \times \mathbb{R}, \\ v_0(0, y) = 0, \quad \partial_\zeta(0, y) \ge 0, \quad & y \in \Gamma. \end{aligned} \tag{2.12}$$

while from (2.11) we get

$$\int_{\mathbb{R}} [\, \frac{1}{2}|\partial_\zeta v_0|^2 + \frac{1}{4}(1 - v_0^2)^2 \,]\, d\zeta < +\infty \tag{2.13}$$

Conditions (2.13) and (2.12) force $v_0(y, \zeta) = w(\zeta)$ where w is the unique solution of the ordinary differential equation

$$w'' + w - w^3 = 0, \quad w(0) = 0, \quad w(\pm\infty) = \pm 1, \tag{2.14}$$

namely

$$w(\zeta) := \tanh(\zeta/\sqrt{2}). \tag{2.15}$$

On the other hand, substitution yields that $v_1(y,\zeta)$ satisfies

$$\partial_\zeta^2 v_1 + (1-3w(\zeta)^2)v_1 \;=\; H_\Gamma(y)\, w'(\zeta), \quad \zeta \in (-\infty,\infty) \tag{2.16}$$

Testing this equation against $w'(\zeta)$ and integrating by parts in ζ we get the relation

$$H_\Gamma(y) = 0 \quad \text{for all} \quad y \in \Gamma$$

which tells us precisely that Γ must be a minimal surface, as expected. Hence, we get $v_1 = -h_0(y)w'(\zeta)$ for a certain function $h_0(y)$. As a conclusion, from (2.10) and a Taylor expansion, we can write

$$v_\varepsilon(y,\zeta) = w(\zeta - \varepsilon h_0(y)) + \varepsilon^2 v_2 + \cdots$$

It is convenient to write this expansion in terms of the variable $t = \zeta - \varepsilon h_0(y)$ in the form

$$v_\varepsilon(y,\zeta) = w(t) + \varepsilon^2 v_2(t,y) + \varepsilon^3 v_3(t,y) + \cdots \tag{2.17}$$

Using expression (2.8) and the fact that Γ is a minimal surface, we expand

$$H_{\Gamma^{\varepsilon\zeta}}(y) = \varepsilon^2\zeta\,|A_\Gamma(y)|^2 + \varepsilon^3\zeta^2\, H_3(y) + \cdots$$

where

$$|A_\Gamma|^2 = \sum_{i=1}^{8} k_i^2, \quad H_3 = \sum_{i=1}^{8} k_i^3.$$

Thus setting $t = \zeta - \varepsilon h_0(y)$ and using (2.17), we compute

$$\begin{aligned} 0 = \Delta v_\varepsilon + v_\varepsilon + v_\varepsilon^3 = \; & [\partial_t^2 + (1-3w(t)^2)]\,(\varepsilon^2 v_2 + \varepsilon^3 v_3) \\ -w'(t)\,[\,\varepsilon^3\Delta_\Gamma h_0 \; & + \varepsilon^3 H_3\, t^2 + \varepsilon^2 |A_\Gamma|^2\,(t+\varepsilon h_0)\,] + \; O(\varepsilon^4). \end{aligned}$$

And then letting $\varepsilon \to 0$ we arrive to the equations

$$\partial_t^2 v_2 + (1-3w^2)v_2 = |A_\Gamma|^2\, tw', \tag{2.18}$$

$$\partial_t^2 v_3 + (1-3w^2)v_3 = [\Delta_\Gamma h_0 + |A_\Gamma|^2 h_0 + H_3\, t^2]\, w'. \tag{2.19}$$

Equation (2.18) has a bounded solution since $\int_{\mathbb{R}} tw'(t)^2\,dt = 0$ Instead the bounded solvability of (2.19) is obtained if and only if h_0 solves the following elliptic equation in Γ.

$$\mathcal{J}_\Gamma[h_0](y) \;:=\; \Delta_\Gamma h_0 + |A_\Gamma|^2 h_0 \;=\; c\sum_{i=1}^{8} k_i^3 \quad \text{in } \Gamma, \tag{2.20}$$

where $c = -\int_{\mathbb{R}} t^2 {w'}^2\,dt / \int_{\mathbb{R}} {w'}^2\,dt$. $\mathcal{J}_\Gamma$ is by definition the *Jacobi operator* of the minimal surface Γ.

We deal with the problem of constructing entire solutions of Equation (2.2), that exhibit the asymptotic behavior described above, around a given, fixed

minimal hypersurface Γ that splits the space $\mathbb{R}^N$ into two components, and for which the coordinates (2.6) are defined for some uniform $\delta > 0$. A key element for such a construction is the precisely the question of solvability of Equation (2.20), that determines at main order the deviation of the nodal set of the solution from Γ.

In terms of the original problem (2.1), the issue is to consider a large dilation of Γ,

$$\Gamma_\varepsilon \ := \ \varepsilon^{-1}\Gamma,$$

and find an entire solution u_ε to problem (2.1) such that for a function h_ε defined on Γ with

$$\sup_{\varepsilon>0} \|h_\varepsilon\|_{L^\infty(\Gamma)} \ < \ +\infty, \tag{2.21}$$

we have

$$u_\varepsilon(x) = w(\zeta - \varepsilon h_\varepsilon(\varepsilon y)) + O(\varepsilon^2), \tag{2.22}$$

uniformly for

$$x = y + \zeta\nu(\varepsilon y), \quad |\zeta| \le \frac{\delta}{\varepsilon}, \quad y \in \Gamma_\varepsilon,$$

while

$$|u_\varepsilon(x)| \to 1 \quad \text{as } \operatorname{dist}(x, \Gamma_\varepsilon) \to +\infty. \tag{2.23}$$

We shall answer affirmatively this question in some important examples for Γ. One is a nontrivial minimal graph in $\mathbb{R}^9$. The solution found provides a negative answer to to a famous question due to Ennio De Giorgi [25]. On the other hand, in $\mathbb{R}^3$ we find a broad new class of entire solutions with finite Morse index, which suggests analogs of De Giorgi's conjecture for solutions of (2.1) in parallel with known classification results for minimal surfaces.

3. From Bernstein's to De Giorgi's Conjecture

Ennio De Giorgi [25] formulated in 1978 the following celebrated conjecture concerning entire solutions of equation (2.1).

De Giorgi's Conjecture: *Let u be a bounded solution of equation* (2.1) *such that $\partial_{x_N} u > 0$. Then the level sets $[u = \lambda]$ are all hyperplanes,* **at least** *for dimension $N \le 8$.*

Equivalently, u must depend only on one Euclidean variable so that it must have the form $u(x) = w((x-p)\cdot\nu)$ for some $p \in \mathbb{R}^N$ and some ν with $|\nu| = 1$ and $\nu_N > 0$.

The condition $\partial_{x_N} u > 0$ implies that the level sets of u are all graphs of functions of the first $N-1$ variables. As we have discussed in the previous section, level sets of solutions with a transition are closely connected to minimal hypersurfaces. De Giorgi's conjecture is in fact a parallel to the following classical statement.

Bernstein's conjecture: *A minimal hypersurface in $\mathbb{R}^N$, which is also the graph of a smooth entire function of $N-1$ variables, must be a hyperplane.*

In other words, if Γ is an *entire minimal graph*, namely

$$\Gamma = \{(x', x_N) \mid x' \in \mathbb{R}^{N-1},\ x_N = F(x')\} \tag{3.1}$$

where F solves the minimal surface equation

$$H_\Gamma \equiv \nabla \cdot \left(\frac{\nabla F}{\sqrt{1+|\nabla F|^2}} \right) = 0 \quad \text{in } \mathbb{R}^{N-1}, \tag{3.2}$$

then Γ must be a hyperplane, hence F must be a linear affine function.

Bernstein's conjecture is known to be true up to dimension $N = 8$, see Simons [80] and references therein, while it is *false* for $N \geq 9$, as proven by Bombieri, De Giorgi and Giusti [12], by building a nontrivial solution to Equation (3.2). Let us write $x' \in \mathbb{R}^8$ as $x' = (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^4 \times \mathbb{R}^4$. Let us consider the set

$$T \ := \ \{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^8 \mid \ |\mathbf{v}| > |\mathbf{u}| > 0 \ \}. \tag{3.3}$$

The set $\{|\mathbf{u}| = |\mathbf{v}|\} \in \mathbb{R}^8$ is Simons' minimal cone [80]. The solution found in [12] is radially symmetric in both variables, namely $F = F(|\mathbf{u}|, |\mathbf{v}|)$. In addition, F is positive in T and it vanishes along Simons' cone. Moreover, it satisfies

$$F(|\mathbf{u}|, |\mathbf{v}|) = -F(|\mathbf{v}|, |\mathbf{u}|) \ . \tag{3.4}$$

Let us write $(|\mathbf{u}|, |\mathbf{v}|) = (r\cos\theta, r\sin\theta)$. In [30] it is found that there is a function $g(\theta)$ with

$$g(\theta) > 0 \quad \text{in } (\pi/4, \pi/2), \quad g'(\pi/2) = 0 = g(\pi/4), \quad g'(\pi/4) > 0,$$

such that for some $\sigma > 0$,

$$F(|\mathbf{u}|, |\mathbf{v}|) \ = \ g(\theta)\, r^3 + O(r^{-\sigma}) \quad \text{in } T. \tag{3.5}$$

De Giorgi's conjecture has been established in dimensions $N = 2$ by Ghoussoub and Gui [41] and for $N = 3$ by Ambrosio and Cabré [15]. Savin [76] proved its validity for $4 \leq N \leq 8$ under the additional assumption

$$\lim_{x_N \to \pm\infty} u(x', x_N) = \pm 1 \quad \text{for all} \quad x' \in \mathbb{R}^{N-1}. \tag{3.6}$$

Farina and Valdinoci [38] replaced condition (3.6) by the less restrictive assumption that the profiles at infinity are two-dimensional functions, or that their level sets are complete graphs. Condition (3.6) is related to the so-called Gibbons' Conjecture:

Gibbons' Conjecture: Let u be a bounded solution of equation (2.1) *satisfying Condition (3.6) uniformly in x'. Then the level sets of u are all hyperplanes.*

Gibbons' Conjecture has been established in all dimensions with different methods by Caffarelli and Córdoba [17], Farina [36], Barlow, Bass and Gui [10], and Berestycki, Hamel, and Monneau [11]. In [17, 10] it is proven that the conjecture is true for any solution that has one level set which is a globally Lipschitz graph.

The following result disproves De Giorgi's statement for $n \geq 9$.

Theorem 1 ([30, 31]). *Let $N \geq 9$. Then there is an entire minimal graph Γ which is not a hyperplane, such that all $\varepsilon > 0$ sufficiently small there exists a bounded solution $u_\varepsilon(x)$ of equation* (2.1) *that satisfies properties* (2.21)-(2.23). *Besides, $\partial_{x_N} u_\varepsilon > 0$ and u_ε satisfies condition* (3.6).

A counterexample to De Giorgi's conjecture in dimension $N \geq 9$ was believed to exist for a long time. Partial progress in this direction was made by Jerison and Monneau [51] and by Cabré and Terra [13]. See also the survey article by Farina and Valdinoci [37].

3.1. Outline of the proof. For a small $\varepsilon > 0$ we look for a solution u_ε of the form (near Γ_ε),

$$u_\varepsilon(x) = w(\zeta - \varepsilon h(\varepsilon y)) + \phi(\zeta - \varepsilon h(\varepsilon y), y), \quad x = y + \zeta \nu(\varepsilon y) \tag{3.7}$$

where $y \in \Gamma_\varepsilon$, ν designates a unit normal to Γ with $\nu_N > 0$, h is a function defined on Γ, which is left as a parameter to be adjusted. Setting $r(y', y_9) = |y'|$, we assume a priori in h that

$$\|(1+r^2) D_\Gamma h\|_{L^\infty(\Gamma)} + \|(1+r)h\|_{L^\infty(\Gamma)} \ \leq \ M \tag{3.8}$$

for some large, fixed number M, also with a uniform control on $(1+r^3) D^2_\Gamma h$.

Letting $f(u) = u - u^3$ and using Expression (2.7) for the Laplacian, the equation becomes

$$\begin{aligned} S(u_\varepsilon) := \Delta u_\varepsilon + f(u_\varepsilon) \ = \ & \\ \Delta_{\Gamma^\zeta_\varepsilon} u_\varepsilon - \varepsilon H_{\Gamma^\zeta_\varepsilon}(\varepsilon y)\, \partial_\zeta u_\varepsilon \ + \ & \\ \partial^2_\zeta u_\varepsilon + f(u_\varepsilon) \ = 0, \quad & y \in \Gamma_\varepsilon, \ |\zeta| < \delta/\varepsilon. \end{aligned} \tag{3.9}$$

Letting $t = \zeta - \varepsilon h(\varepsilon y)$, we look for u_ε of the form

$$u_\varepsilon(t, y) = w(t) + \phi(t, y)$$

for a small function ϕ. The equation in terms of ϕ becomes

$$\partial^2_t \phi + \Delta_{\Gamma_\varepsilon} \phi + B\phi + f'(w(t))\phi + N(\phi) + E \ = \ 0. \tag{3.10}$$

where B is a small linear second order operator, and

$$E = S(w(t)), \quad N(\phi) = f(w+\phi) - f(w) - f'(w)\phi \approx f''(w)\phi^2.$$

While the expression (3.10) makes sense only for $|t| < \delta\varepsilon^{-1}$, it turns out that the equation in the entire space can be reduced to one similar to (3.10) in entire $\mathbb{R} \times \Gamma_\varepsilon$, where E and the undefined coefficients in B are just cut-off far away, while the operator N is slightly modified by the addition of a small nonlinear, nonlocal operator of ϕ. Rather than solving this problem directly we carry out an infinite dimensional form of Lyapunov-Schmidt reduction, considering a projected version of it,

$$\begin{aligned} \partial_t^2\phi + \Delta_{\Gamma_\varepsilon}\phi + B\phi + f'(w(t))\phi + N(\phi) + E \;=\; & \\ c(y)w'(t) \quad \text{in } \mathbb{R}\times\Gamma_\varepsilon, & \\ \int_{\mathbb{R}} \phi(t,y)w'(t)\,dt \;=\; 0 \quad \text{for all} \quad y \in \Gamma_\varepsilon. & \end{aligned} \tag{3.11}$$

the error of approximation E has roughly speaking a bound $O(\varepsilon^2 r(\varepsilon y)^{-2} e^{-\sigma|t|})$, and it turns out that one can find a solution $\phi = \Phi(h)$ to problem (3.11) with the same bound. We then get a solution to our original problem if h is such that $c(y) \equiv 0$. Thus the problem is reduced to finding h such that

$$c(y)\int_{\mathbb{R}} w'^2 = \int_{\mathbb{R}} (E + B\Phi(h) + N(\Phi(h)))\, w' dt \equiv 0.$$

A computation similar to that in the formal derivation yields that this problem is equivalent to a small perturbation of Equation (2.20)

$$\mathcal{J}_\Gamma(h) := \Delta_\Gamma h + |A_\Gamma|^2 h \;=\; c\sum_{i=1}^{8} k_i^3 + \mathcal{N}(h) \quad \text{in } \Gamma, \tag{3.12}$$

where $\mathcal{N}(h)$ is a small operator. From an estimate by Simon [79] we know that $k_i = O(r^{-1})$. Hence $H_3 := \sum_{i=1}^8 k_i^3 = O(r^{-3})$. A central point is to show that the unperturbed equation (2.20) has a solution $h = O(r^{-1})$, which justifies a posteriori the assumption (3.8) made originally on h. This step uses the asymptotic expression (3.5). The symmetries of the surface allow to reduce the problem to solving it in T with zero Dirichlet boundary conditions on Simons' cone. We have that $H_3 = O(g(\theta)r^{-3})$ and one gets a positive barrier of size $O(r^{-1})$. The operator $\mathcal{J}_\Gamma$ satisfies maximum principle and existence thus follows. The full nonlinear equation is then solved with the aid of contraction mapping principle. The detailed proof of this theorem is contained in [30].

The program towards the counterexample in [51] and [15] is based on an analogous one in Bernstein's conjecture: the existence of the counterexample is reduced to establishing the minimizing character of a *saddle solution* in $\mathbb{R}^8$ that vanishes on Simon's cone. Our approach of direct construction is actually applicable to build unstable solutions associated to general minimal surfaces, as we illustrate in the next section. We should mention that method of infinite dimensional reduction for the Allen Cahn equation in compact settings has precedents with similar flavor in [73], [55], [29]. Using variational approach, local minimizers were built in [54].

4. Finite Morse Index Solutions of the Allen-Cahn Equation in $\mathbb{R}^3$

The assumption of monotonicity in one direction for the solution u in De Giorgi's conjecture implies a form of stability, locally minimizing character for u when compactly supported perturbations are considered in the energy. Indeed, the linearized operator $L = \Delta + (1 - 3u^2)$, satisfies maximum principle since $L(Z) = 0$ for $Z = \partial_{x_N} u > 0$. This implies stability of u, in the sense that its associated quadratic form, namely the second variation of the corresponding energy,

$$\mathcal{Q}(\psi, \psi) := \int_{\mathbb{R}^3} |\nabla \psi|^2 + (3u^2 - 1)\, \psi^2 \tag{4.1}$$

satisfies $\mathcal{Q}(\psi, \psi) > 0$ for all $\psi \neq 0$ smooth and compactly supported. Stability of u is indeed sufficient for De Giorgi's statement to hold in dimension $N = 2$, as observed by Dancer [22]. This question is open for $3 \leq N \leq 8$. The monotonicity assumption actually implies the globally minimizing character of the solution on each compact set, subject to its own boundary conditions, see [1].

The *Morse index* $m(u)$ is defined as the maximal dimension of a vector space E of compactly supported functions such that

$$\mathcal{Q}(\psi, \psi) < 0 \quad \text{for all} \quad \psi \in E \setminus \{0\}.$$

In view of the discussion so far, it seems natural to associate complete, embedded minimal surfaces Γ with finite Morse index, and solutions of (2.1). The *Morse index* of the minimal surface Γ, $i(\Gamma)$, has a similar definition relative to the quadratic form for its Jacobi operator $\mathcal{J}_\Gamma := \Delta_\Gamma + |A_\Gamma|^2$: The number $i(\Gamma)$ is the largest dimension for a vector spaced E of compactly supported smooth functions in Γ with

$$\int_\Gamma |\nabla k|^2\, dV \; - \; \int_\Gamma |A|^2 k^2\, dV \; < 0 \quad \text{for all} \quad k \in E \setminus \{0\}.$$

We point out that for complete, embedded surfaces, finite index is equivalent to *finite total curvature,* namely

$$\int_\Gamma |K|\, dV \; < \; +\infty$$

where K denotes Gauss curvature of the manifold, see §7 of [48] and references therein.

4.1. Embedded minimal surfaces of finite total curvature.

The theory of embedded, minimal surfaces of finite total curvature in $\mathbb{R}^3$, has reached a notable development in the last 25 years. For more than a century, only two examples of such surfaces were known: the plane and the catenoid.

The first nontrivial example was found in 1981 by C. Costa, [19, 20]. The *Costa surface* is a genus one minimal surface, complete and properly embedded, which outside a large ball has exactly three components (its *ends*), two of which are asymptotically catenoids with the same axis and opposite directions, the third one asymptotic to a plane perpendicular to that axis. The complete proof of embeddedness is due to Hoffman and Meeks [49]. In [50] these authors generalized notably Costa's example by exhibiting a class of three-end, embedded minimal surface, with the same look as Costa's far away, but with an array of tunnels that provides arbitrary genus $\ell \geq 1$. This is known as the Costa-Hoffman-Meeks surface with genus ℓ.

As a special case of the main results of [32] we have the following

Theorem 2 ([32]). *Let $\Gamma \subset \mathbb{R}^3$ be either a catenoid or a Costa-Hoffman-Meeks surface with genus $\ell \geq 1$. Then for all sufficiently small $\varepsilon > 0$ there exists a solution u_ε of Problem* (2.1) *with the properties* (2.21)-(2.23). *In the case of the catenoid, the solution found is radially symmetric in two of its variables and $m(u_\varepsilon) = 1$. For the Costa-Hoffman-Meeks surface with genus $\ell \geq 1$, we have $m(u_\varepsilon) = 2\ell + 3$.*

4.2. A general statement. In what follows Γ designates a complete, embedded minimal surface in $\mathbb{R}^3$ with finite total curvature. Then Γ is orientable and the set $\mathbb{R}^3 \setminus \Gamma$ has exactly two components S_+, S_-, see [48]. In what follows we fix a continuous choice of unit normal field $\nu(y)$, which conventionally we take it to point towards S_+.

For $x = (x', x_3) \in \mathbb{R}^3$, we denote as before, $r = r(x) = |x'|$. It is known that after a suitable rotation of the coordinate axes, outside the infinite cylinder $r < R_0$ with sufficiently large radius R_0, Γ decomposes into a finite number m of unbounded components $\Gamma_1, \ldots, \Gamma_m$, its *ends*. From a result in [78], we know that asymptotically each end of Γ_k either resembles a plane or a catenoid. More precisely, Γ_k can be represented as the graph of a function F_k of the first two variables,

$$\Gamma_k = \{\, y \in \mathbb{R}^3 \;/\; r(y) > R_0,\ y_3 = F_k(y') \,\}$$

where F_k is a smooth function which can be expanded as

$$F_k(y') = a_k \log r + b_k + b_{ik} \frac{y_i}{r^2} + O(r^{-3}) \quad \text{as } r \to +\infty, \tag{4.2}$$

for certain constants a_k, b_k, b_{ik}, and this relation can also be differentiated. Here

$$a_1 \leq a_2 \leq \ldots \leq a_m\ , \qquad \sum_{k=1}^{m} a_k \ = \ 0\,. \tag{4.3}$$

We say that Γ has *non-parallel ends* if all the above inequalities are strict.

Let us consider the Jacobi operator of Γ

$$\mathcal{J}_\Gamma(h) \ := \ \Delta_\Gamma h + |A_\Gamma|^2 h \tag{4.4}$$

where $|A_\Gamma|^2 = k_1^2 + k_2^2 = -2K$. A smooth function $z(y)$ defined on Γ is called a *Jacobi field* if $\mathcal{J}_\Gamma(z) = 0$. Rigid motions of the surface induce naturally some bounded Jacobi fields: Associated to respectively translations along coordinates axes and rotation around the x_3-axis, are the functions

$$z_1(y) = \nu(y) \cdot e_i, \quad y \in \Gamma, \quad i = 1, 2, 3,$$

$$z_4(y) = (-y_2, y_1, 0) \cdot \nu(y), \quad y \in \Gamma. \tag{4.5}$$

We assume that Γ is *non-degenerate* in the sense that these functions are actually *all* the bounded Jacobi fields, namely

$$\{\ z \in L^\infty(\Gamma)\ /\ \mathcal{J}_\Gamma(z) = 0\ \} \ =\ \text{span}\,\{\ z_1, z_2, z_3, z_4\ \}\,. \tag{4.6}$$

This property is known in some important cases, most notably the catenoid and the Costa-Hoffmann-Meeks surface of any order $\ell \geq 1$. See Nayatani [67, 68] and Morabito [65].

Theorem 3 ([32]). *Let $N = 3$ and Γ be a minimal surface embedded, complete with finite total curvature and non-parallel ends, which is in addition nondegenerate. Then for all sufficiently small $\varepsilon > 0$ there exists a solution u_ε of Problem* (2.1) *with the properties* (2.21)-(2.23). *Moreover, we have*

$$m(u_\varepsilon) \ =\ i(\Gamma).$$

Besides, the solution is non-degenerate, in the sense that any bounded solution of

$$\Delta\phi \,+\, (1 - 3u_\varepsilon^2)\,\phi \,=\, 0 \qquad \text{in } \mathbb{R}^3$$

must be a linear combination of the functions $Z_i,\ i = 1, 2, 3, 4$ defined as

$$Z_i = \partial_i u_\varepsilon, \quad i = 1, 2, 3, \quad Z_4 = -x_2\partial_1 u_\varepsilon + x_1\partial_2 u_\varepsilon.$$

It is well-known that if Γ is a catenoid then $i(\Gamma) = 1$. Moreover, in the Costa-Hoffmann-Meeks surface it is known that $i(\Gamma) = 2\ell + 3$ where ℓ is the genus of Γ. See [67, 68, 65].

4.3. Further comments. In analogy with De Giorgi's conjecture, it seems plausible that qualitative properties of embedded minimal surfaces with finite Morse index should hold for the level sets of finite Morse index solutions of Equation (2.1), provided that these sets are embedded manifolds outside a compact set. As a sample, one may ask for the validity of the following two statements:

- *The level sets of any finite Morse index solution u of* (2.1) *in $\mathbb{R}^3$, such that $\nabla u \neq 0$ outside a compact set should have a finite, even number of catenoidal or planar ends with a common axis.*

The above fact does hold for minimal surfaces with finite total curvature and embedded ends as established by Ossermann and Schoen. On the other hand, the above statement should not hold true if the condition $\nabla u \neq 0$ outside a large ball is violated. For instance, let us consider the octant $\{x_1, x_2, x_3 \geq 0\}$. Problem (2.1) in the octant with zero boundary data can be solved by a super-subsolution scheme (similar to that in [23]) yielding a positive solution. Extending by successive odd reflections to the remaining octants, one generates an entire solution (likely to have finite Morse index), whose zero level set does not have the characteristics above: the condition $\nabla u \neq 0$ far away corresponds to *embeddedness of the ends of the level sets.*

An analog of De Giorgi's conjecture for the solutions that follow in complexity the stable ones, namely those with Morse index one, may be the following:

- *A bounded solution u of (2.1) in $\mathbb{R}^3$ with $i(u) = 1$, and $\nabla u \neq 0$ outside a bounded set, must be axially symmetric, namely radially symmetric in two variables.*

The solution we found, with transition on a dilated catenoid has this property. This statement would be in correspondence with results by Schoen [78] and López and Ros [58]: if $i(\Gamma) = 1$ and Γ has embedded ends, then it must be a catenoid.

5. The Allen-Cahn Equation in $\mathbb{R}^2$

5.1. Solutions with multiply connected nodal set. The only minimal surface Γ that we can consider in this case is a straight line, to which the trivial solution depending on its normal variable can be associated.

A class of solutions to (2.1) with a *finite number of transition lines*, likely to have finite Morse index, has been recently built in [34]. The location and shape of these lines is governed by the *Toda system*, a classical integrable model for scattering of particles on a line under the action of a repulsive exponential potential:

$$\frac{\sqrt{2}}{24} f_j'' = e^{-\sqrt{2}(f_j - f_{j-1})} - e^{-\sqrt{2}(f_{j+1} - f_j)}, \quad j = 1, \dots k, \tag{5.1}$$

$f_0 \equiv -\infty$, $f_{k+1} \equiv +\infty$. It is known that for a given solution there exist numbers $a_j^{\pm}, b_j^{\pm}$ such that

$$f_j(z) = a_j^{\pm}|z| + b_j^{\pm} + O(e^{-|z|}) \quad \text{as } z \to \pm\infty$$

where $a_j^{\pm} < a_{j+1}^{\pm}$, $j = 1, \dots, k-1$ (long-time scattering).

The role of this system in the construction of solutions with multiple transition lines in the Allen-Cahn equation in bounded domains was discovered in [29]. In entire space the following result holds.

Theorem 4 ([34]). *Given a solution f of (5.1) if we scale*

$$f_{\varepsilon,j}(z) := \sqrt{2}\,(j - \frac{k+1}{2}) \log \frac{1}{\varepsilon} + f_j(\varepsilon z),$$

then for all small ε there is a solution u_ε with k transitions layers near the lines $x_2 = f_{\varepsilon,j}(x_1)$. More precisely,

$$u_\varepsilon(x_1, x_2) = \sum_{j=1}^{k} (-1)^{j-1} w(x_1 - f_{\varepsilon,j}(x_2)\,) + \sigma_k + O(\varepsilon), \tag{5.2}$$

where $\sigma_k = -\frac{1}{2}(1 + (-1)^k)$.

The transition lines are therefore nearly parallel and asymptotically straight. In particular, if $k = 2$ and f solves the ODE

$$\frac{\sqrt{2}}{24} f''(z) = e^{-2\sqrt{2} f(z)}, \quad f'(0) = 0,$$

and $f_\varepsilon(z) := \sqrt{2} \log \frac{1}{\varepsilon} + f(\varepsilon z)$, then there exists a solution u_ε to (2.1) in $\mathbb{R}^2$ with

$$u_\varepsilon(x_1, x_2) = w(x_1 + f_\varepsilon(x_2)\,) + w(x_1 - f_\varepsilon(x_2)) \; - 1 + O(\varepsilon). \tag{5.3}$$

The formal reason for the appearance of the Toda system can be explained as follows: Let us consider the function

$$u_*(x_1, x_2) = \sum_{j=1}^{k} (-1)^{j-1} w(x_1 - f_j(x_2)\,) + \sigma_k$$

and assume that the f_j's are ordered and very distant one to each other. Then the energy

$$J_S(u_*) = \frac{1}{2} \int_S |\partial_{x_2} u_*|^2 + |\partial_{x_1} u_*|^2 + \frac{1}{4} \int_S (1 - u_*^2)^2$$

computed in a finite strip $S = \mathbb{R} \times (-\ell, \ell)$ becomes at main order, after some computation,

$$J_S(u_*) \approx 2\ell\,[\frac{1}{2} \int_{\mathbb{R}} |w'|^2 + \frac{1}{4} \int_{\mathbb{R}} (1 - w^2)^2\,] + c_1 \sum_{j=1}^{k} \int_{-\ell}^{\ell} |f_j'|^2 - c_2 \sum_{i \neq j} \int_{-\ell}^{\ell} e^{-\sqrt{2}|f_i - f_j|}$$

for certain explicit constants c_1 and c_2. Assuming that the quantities $e^{-\sqrt{2}|f_i - f_j|}$ are negligible for $|i - j| \geq 2$, we obtain for the approximate equilibrium condition of the functions f_j, precisely the system (5.1).

5.2. Remarks. The solutions (5.2) show a major difference between the minimal surface problem and the Allen-Cahn equation, as it is the fact that two separate interfaces *interact*, leading to a major deformation in their asymptotic shapes. We believe that these examples should be prototypical of bounded finite Morse index solutions of (2.1). A finite Morse index solution u is stable outside a bounded set. If we follow a component of its nodal set along a unbounded sequence, translation and a standard compactness argument leads in the limit to a stable solution. Hence from the result in [22] its profile must be one-dimensional and hence its nodal set is a straight line. This makes it plausible that the ends of the nodal set of u *are asymptotically a finite, even number of straight lines*. If this is the case, those lines are not disposed in arbitrary way: Gui [46] proved that if $e_1,\dots e_{2k}$ are unit vectors in the direction of the ends of the nodal set of a solution of (2.1) in $\mathbb{R}^2$, then the balancing formula $\sum_{j=1}^{2k} e_j = 0$ holds.

As we have mentioned, another finite Morse solution is known, [23], the so-called saddle solution. It is built by positive barriers with zero boundary data on a quadrant, and then extended by odd reflections to the rest of the plane, so that its nodal set is an infinite cross, hence having 4 straight ends. The saddle solution has Morse index 1, see [77]. This is also formally the case for the solutions (5.3), which also has 4 ends.

An interesting question is whether the parameter ε of the solutions (5.3) can be continued to increase the nearly zero angle between ends up to $\frac{\pi}{2}$, the case of the saddle solution. Similarly, a saddle solutions with $2k$ ends with consecutive angles $\frac{\pi}{k}$ has been built in [2]. One may similarly ask whether this solution is in some way connected to the $2k$-end family (5.2).

6. The Stationary NLS and the Yamabe Equations

6.1. The standing wave problem for NLS. We shall discuss some results on the problem

$$\Delta u + |u|^{p-1}u - u = 0 \quad \text{in } \mathbb{R}^N \tag{6.1}$$

where $p > 1$. Equation (6.1) arises for instance as the standing-wave problem for the standard nonlinear Schrödinger equation

$$i\psi_t = \Delta\psi + |\psi|^{p-1}\psi, \tag{6.2}$$

corresponding to that of solutions of the form $\psi(y,t) = u(y)e^{-it}$. It also arises in nonlinear models in Turing's theory biological theory of pattern formation, such as the Gray-Scott or Gierer-Meinhardt systems, [44, 43]. The positive solutions of (6.1) which decay to zero at infinity are well understood. Problem

(6.1) has a radially symmetric solution $w_N(y)$ which approaches 0 at infinity provided that

$$1 < p < \begin{cases} \frac{N+2}{N-2} & \text{if } N \geq 3, \\ +\infty & \text{if } N = 1, 2, \end{cases}$$

see [81, 7]. This solution is unique [56], and actually any positive solution to (6.1) which vanishes at infinity must be radially symmetric around some point [42].

Variations of Problem (6.1), where the space homogeneity is broken by the action of an external potential or boundary conditions in a domain, have been broadly treated in the PDE literature in the last two decades, especially concerning the construction of *positive solutions*. Widely studied has been for instance a singular perturbation problem of the form

$$\varepsilon^2 \Delta - V(x)u + |u|^{p-1}u = 0 \tag{6.3}$$

where ε is a small parameter, or in a bounded domain with $V \equiv 1$, under Dirichlet or Neumann boundary conditions. Many constructions in the literature refer to "multi-bump solutions", built from a perturbation of the superposition of suitably scaled copies of the basic radial bump w_N. The location of their maxima is determined typically by a criterion related either with the potential or the geometry of the underlying domain. Among other contributions, we refer the reader to the works [4, 6, 26, 27, 57, 39, 45, 52, 69, 70, 71, 24, 75, 83] and their references. Solutions concentrating on a higher dimensional sets have been considered for instance in [60, 61, 28, 59].

It is natural to ask about positive solutions to (6.1) which do not vanish at infinity.

For instance, let us consider the solution $w := w_1$ of (6.1) in $\mathbb{R}$,

$$\begin{aligned} w'' - w + w^p = 0, \quad w > 0, \quad &\text{in } \mathbb{R}, \\ w'(0) = 0, \quad w(\pm\infty) = 0. \end{aligned} \tag{6.4}$$

Then the functions $u(x, z) := w(x - a)$, $a \in \mathbb{R}$, define a class of positive solutions on (6.1) in $\mathbb{R}^2$, which vanish in all but one space direction, corresponding to single "bump lines", very much in analogy to the trivial single transition solutions to the Allen-Cahn equation induced by (2.14). In [8], these solutions of (6.1) were found to be isolated in a uniform topology which avoids oscillations at infinity. In constrast, in [21] it is found that a there is continuum of solutions $w_\delta(x, z)$ which are periodic in z and decay exponentially in x, bifurcating from $w(x)$.

A big qualitative difference between the homoclinic solution (6.4) and the heteroclinic solution (2.14) is that the latter is stable, and that avoids these bifurcations. Instead, there is a positive eigenvalue λ_1 to with positive eigenfunction to the linearized equation

$$Z'' + (pw^{p-1} - 1)Z - \lambda_1 Z = 0 \quad \text{in } \mathbb{R}, \quad Z(\pm\infty) = 0,$$

and the bifurcating *Dancer solutions* can be expanded as .

$$w^\delta(x,z) = w(x) + \delta Z(x)\cos(\sqrt{\lambda_1}z) + O(\delta^2 e^{-|x|}). \tag{6.5}$$

Intuitively, as δ increases, the period becomes long and the oscillating amplitude largely varies: in fact a simple variational argument using symmetries gives also the existence of a solution $w^T(x,z)$ with a large period $T \gg 1$ whose profile is an "infinite bump array" solution like

$$w^T(x,z) \approx \sum_{k=-\infty}^{\infty} w_2(x, z - k\,T), \tag{6.6}$$

where w_2 is the radial positive solution that decays to zero of (6.1). The solutions to (6.6)

Independently in [33] and [62], positive solutions that glue together respectively bump-lines and infinite bump arrays have been built.

The result in [33] is the exact analog of Theorem 4, now with a Toda system of the form

$$c_p f_j'' = e^{-(f_j - f_{j-1})} - e^{-(f_{j+1} - f_j)}, \quad j = 1, \dots k, \tag{6.7}$$

$f_0 \equiv -\infty,\ f_{k+1} \equiv +\infty$, where c_p is a explicit positive constant.

Theorem 5 ([33]). *Given a solution f of* (6.7) *if we scale*

$$f_{\varepsilon,j}(z) := \sqrt{2}\,(j - \frac{k+1}{2})\log\frac{1}{\varepsilon} + f_j(\varepsilon z),$$

then for all small ε there is a positive solution u_ε of (6.1) with k bump lines:

$$u_\varepsilon(x,z) = \sum_{j=1}^{k} w(x - f_{\varepsilon,j}(z)\,) + O(\varepsilon). \tag{6.8}$$

The profile of the solution (6.8) can actually be more accurately described as a superposition of bifurcating Dancer solutions (6.5) w^{δ_j}, with respective axes given at main order by the straight line asymptote of to the graphs of the f_j's, and with $\delta_j(\varepsilon) \to 0$, plus a remainder that decays away and along these lines.

In [62] a solution was built close to a given finite number of halves of infinite bump arrays (6.6), with sufficiently large T, emanating from the origin, and along three divergent rays with sufficiently large mutual angles. The solutions in [33] and those in [62] may belong to endpoints of families with opposite size in their Dancer parameters, in a way perhaps similar as the solutions in Theorem 4 are expected to connect to the symmetric saddle solutions, but this is still far from understood. Obtaining (even partial) classification of the positive solutions of (6.1) is presumably much harder than in the Allen-Cahn equation.

In particular, Morse index of the solutions built turn out to be infinite due to the oscillations along their ends.

Another interesting issue is that of understanding *sign changing solutions*. Even those with finite energy (and finite morse index) can exhibit very complex patterns. From Ljusternik-Schirelmann theory applied to the energy functional

$$J(u) = \frac{1}{2}\int_{\mathbb{R}^N} |\nabla u|^2 + u^2 - \frac{1}{p+1}\int_{\mathbb{R}^N} |u|^{p+1}$$

it is known that (6.1) possesses and infinite number of radially symmetric solutions. Nonradial solutions have been built in [66], as a small perturbation of a configuration of half arrays (6.6) symmetrically disposed, cut-off far away outside a disk of very large radius, and so that the sides of the regular polygon thus formed is filled with alternating sign, nearly equidistant bumps w_2.

6.2. The Yamabe equation in $\mathbb{R}^N$. Let us consider the equation at the critical exponent

$$\Delta u + |u|^{\frac{4}{N-2}} u = 0 \quad \text{in } \mathbb{R}^N \tag{6.9}$$

$N \geq 3$. It is known that a positive solution to this problem must be equal to one of the Aubin-Talenti extremals for Sobolev's embedding,

$$w_{\mu,\xi}(x) = \alpha_N \left(\frac{\mu}{\mu^2 + |x-\xi|^2}\right)^{\frac{N-2}{2}}, \quad \alpha_N = (N(N-2))^{\frac{N-2}{4}}. \tag{6.10}$$

See [72, 5, 82, 18].

The energy associated to Problem (6.9) is given by

$$J(u) = \frac{1}{2}\int_{\mathbb{R}^N} |\nabla u|^2 - \frac{N-2}{2N}\int_{\mathbb{R}^N} |u|^{\frac{2N}{N-2}}.$$

We consider the common value of the energy of the solitons (6.10),

$$S_N := J(w_{\mu,\xi}).$$

Concerning sign changing solutions the whole picture is still far from understood. To our knowledge, only one result is available. Ding [40] proved the existence of infinitely many solutions within a class of solutions which, when, after the equation is lifted to the sphere S^N, it is radially symmetric in two variables. The class of such functions turns out to regain the loss of compactness in Sobolev's embedding, and then Ljusternik-Schnirelmann arguments apply. No further information on the solutions is available. Understanding solutions to (6.9) and its energy levels is an major issue in the analysis of blow-up and well-prosedness for the NLS (6.2) at the critical exponent, in a program initiated in [53].

We have the following result, a special case of that in [35], which describes in precise terms a class of finite energy solutions of (6.9) which do not have the radial symmetries in [40].

Let us consider the points

$$\xi_j := (e^{2\pi ij/k}, 0) \in \mathbb{C} \times \mathbb{R}^{N-2} = \mathbb{R}^N, \quad j = 1, \dots, k$$

where

Theorem 6 ([35]). *for any k sufficiently large, there exists a solution u_k of (6.9) with the form*

$$u_k(x) = w_{1,0}(x) - \sum_{j=1}^{k} w_{\mu_j,\xi_j}(x) + o(1)$$

where for a certain number $\mu_N > 0$,

$$\mu_j = \frac{\nu_N}{k^2}$$

and $o(1) \to 0$ uniformly in $\mathbb{R}^N$ as $k \to +\infty$. Besides we have

$$J(u_k) = (k+1)S_N + O(1)$$

as $k \to \infty$.

A characteristic of this problem is the fact that eventually the concentration set becomes higher dimensional, namely a copy of S^1 in $\mathbb{R}^N$, in spite of being the context just discrete. The hidden parameter here it is of course the number of bubbles. This concentration phenomena can be regarded as somehow intermediate between point and continuum concentration. The idea of using the number of concentrating cells as a singular perturbation parameter appears already in the context of critical problems in [84].

The result o Theorem 6 extends considerably to similar patterns where the limiting concentration set, is, after stereographic projection, a submanifold of the sphere S^N with suitable rotation invariances.

Again, when the Yamabe equation is perturbed by space inhomogeneities or by exponents close to critical, many results on construction and classification of bubbling solutions are present in the literature, but we will not survey them here. The analysis of bubbling solutions has been a central tool for instance in the understanding of the Yamabe and prescribed scalar curvature problems. For changing sign solutions of equation (6.9) in dimension $N = 3$, an analysis of the topology of level sets of the associated energy for low energies is present in [9].

References

[1] G. Alberti, L. Ambrosio and X. Cabré, *On a long-standing conjecture of E. De Giorgi: symmetry in 3D for general nonlinearities and a local minimality property*, Acta Appl. Math. 65 (2001) (Special issue dedicated to Antonio Avantaggiati on the occasion of his 70th birthday), no. 1–3, 9–33.

[2] F. Alessio, A. Calamai, and P. Montecchiari, *Saddle-type solutions for a class of semilinear elliptic equations*, Adv. Differential Equations 12 (2007), 361–380.

[3] S.M. Allen and J.W. Cahn *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening* Acta Metall. 27 (1979), 1084–1095.

[4] A. Ambrosetti, A., M. Badiale and S. Cingolani, *Semiclassical states of nonlinear Schrödinger equations*, Arch. Rational Mech. Anal. 140 (1997), 285–300.

[5] T. Aubin, *Problmes isoprimtriques et espaces de Sobolev,* J. Differ. Geometry 11 (1976), 573–598.

[6] J. Byeon, Z.-Q. Wang, *Standing waves with a critical frequency for nonlinear Schrödinger equations*, Arch. Rat. Mech. Anal. 65(2002), 295–316.

[7] H. Berestycki, P.-L. Lions, *Nonlinear scalar field equations. I. Existence of a ground state*, Arch. Rational Mech. Anal. 82 (1983), no. 4, 313–345.

[8] J. Busca, P. Felmer, *Qualitative properties of some bounded positive solutions to scalar field equations*, Calc. Var. Partial Differential Equations 13 (2001), no. 2, 191–211.

[9] A. Bahri, S. Chanilo, *The Difference of Topology at Infinity in Changing-Sign Yamabe Problems on S^3 (the Case of Two Masses).* Comm. Pure Appl. Math. 54 No 4 (2001), 450–478.

[10] M.T. Barlow, R.F. Bass and C. Gui, *The Liouville property and a conjecture of De Giorgi*, Comm. Pure Appl. Math. 53(8)(2000), 1007–1038.

[11] H. Berestycki, F. Hamel and R. Monneau, *One-dimensional symmetry of bounded entire solutions of some elliptic equations*, Duke Math. J. 103(2000), 375–396.

[12] E. Bombieri, E. De Giorgi, E. Giusti, *Minimal cones and the Bernstein problem,* Invent. Math. 7 (1969) 243–268.

[13] X. Cabré, J. Terra *Saddle-shaped solutions of bistable diffusion equations in all of $\mathbb{R}^{2m}$.* J. Eur. Math. Soc. 11, Issue 4 (2009), 819–943.

[14] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system. I. Interfacial energy,* J. Chem. Phys 28, 258 (1958).

[15] L. Ambrosio and X. Cabré, *Entire solutions of semilinear elliptic equations in $\mathbb{R}^3$ and a conjecture of De Giorgi,* Journal Amer. Math. Soc. 13 (2000), 725–739.

[16] L. Caffarelli, A. Córdoba, *Uniform convergence of a singular perturbation problem,* Comm. Pure Appl. Math. XLVII (1995), 1–12.

[17] L. Caffarelli, A. Córdoba, *Phase transitions: uniform regularity of the intermediate layers.* J. Reine Angew. Math. 593 (2006), 209–235.

[18] L. Caffarelli, B. Gidas, J. Spruck,*Asymptotic symmetry and local behaviour of semilinear elliptic equations with critical Sobolev growth*, Comm. Pure Appl. Math. 42 (1989), 271–297

[19] C.J. Costa, *Imersoes minimas en $\mathbb{R}^3$ de genero un e curvatura total finita. PhD thesis*, IMPA, Rio de Janeiro, Brasil (1982).

[20] C.J. Costa, *Example of a complete minimal immersions in $\mathbb{R}^3$ of genus one and three embedded ends*, Bol. Soc. Bras. Mat. 15(1–2)(1984), 47–54.

[21] E.N. Dancer, *New solutions of equations on* $\mathbf{R}^n$, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) 30 no. 3–4, 535–563 (2002).

[22] E. N. Dancer, *Stable and finite Morse index solutions on* $\mathbf{R}^n$ or on bounded domains with small diffusion. Trans. Amer. Math. Soc. 357 (2005), no. 3, 1225–1243.

[23] H. Dang, P.C. Fife, L.A. Peletier, *Saddle solutions of the bistable diffusion equation.* Z. Angew. Math. Phys. 43 (1992), no. 6, 984–998

[24] J. Dávila, M. del Pino, M. Musso, J. Wei. *Standing waves for supercritical nonlinear Schrodinger equations,* J. Differential Equations 236 no. 1 (2007), 164–198.

[25] E. De Giorgi, *Convergence problems for functionals and operators,* Proc. Int. Meeting on Recent Methods in Nonlinear Analysis (Rome, 1978), 131–188, Pitagora, Bologna (1979).

[26] M. del Pino, P. Felmer, *Local mountain passes for semilinear elliptic problems in unbounded domains*, Calc. Var. Partial Differential Equations 4 (1996), 121–137.

[27] M. del Pino, P. Felmer, *Semi-classcal states for nonlinear Schrödinger equations*, J. Funct. Anal. 149 (1997), 245–265.

[28] M. del Pino, M. Kowalczyk, J. Wei, *Concentration on curves for nonlinear Schrödinger equations,* Comm. Pure Appl. Math. 60 (2007), no. 1, 113–146.

[29] M. del Pino, M. Kowalczyk, J. Wei, *The Toda system and clustering interfaces in the Allen-Cahn equation*, Arch. Rational Mech. Anal. 190 (2008), no. 1, 141–187.

[30] M. del Pino, M. Kowalczyk, J. Wei, *On De Giorgi's Conjecture in Dimensions* $N \geq 9$, preprint 2008.

[31] M. del Pino, M. Kowalczyk, J. Wei, *A counterexample to a conjecture by De Giorgi in large dimensions*, Comp. Rend. Mathematique 346 (2008), 23–24, 1261–1266.

[32] M. del Pino, M. Kowalczyk, J. Wei, *Entire Solutions of the Allen-Cahn equation and Complete Embedded Minimal Surfaces of Finite Total Curvature in* $\mathbb{R}^3$. Preprint 2009.

[33] M. del Pino, M. Kowalczyk, F. Pacard, J. Wei, *The Toda system and multiple-end solutions of autonomous planar elliptic problems*, Preprint 2007, Adv. Math., to appear.

[34] M. del Pino, M. Kowalczyk, F. Pacard, J. Wei, *Multiple-end solutions to the Allen-Cahn equation in* $\mathbb{R}^2$, J. Funct. Anal. 258 No 2 (2010) 458–503.

[35] M. del Pino, M. Musso, F. Pacard, A. Pistoia, *Torus action of* S^n *and sign changing solutions for conformally invariant equations.* Preprint 2010.

[36] A. Farina, *Symmetry for solutions of semilinear elliptic equations in* $\mathbb{R}^N$ *and related conjectures*, Ricerche Mat. 48(suppl.) (1999), 129–154.

[37] A. Farina and E. Valdinoci, *The state of art for a conjecture of De Giorgi and related questions.* "Reaction-Diffusion Systems and Viscosity Solutions", World Scientific, 2008.

[38] A. Farina and E. Valdinoci, *1D symmetry for solutions of semilinear and quasilinear elliptic equations*, Trans. Amer. Math. Soc. (in press).

[39] A. Floer, A. Weinstein, *Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential*, J. Funct. Anal. 69 (1986), 397–408.

[40] W.Y. Ding, *On a conformally invariant elliptic equation on* $\mathbb{R}^n$*,* Commun. Math. Phys. 107, (1986) 331–335.

[41] N. Ghoussoub and C. Gui, *On a conjecture of De Giorgi and some related problems,* Math. Ann. 311 (1998), 481–491.

[42] B. Gidas, W. M. Ni and L. Nirenberg, *Symmetry of positive solutions of nonlinear elliptic equations in* R^N, Adv. Math. Suppl. Stud. 7A (1981) 369–402.

[43] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik (Berlin) 12 (1972) 30–39.

[44] P. Gray and S.K. Scott, *Autocatalytic reactions in the isothermal, continuous stirred tank reactor: isolas and other forms of multistability*, Chem. Eng. Sci. 38 (1983), 29–43.

[45] C. Gui, J. Wei, *Multiple interior peak solutions for some singular perturbation problems*, J. Differential Equations 158 (1999) 1–27.

[46] C. Gui, *Hamiltonian identities for elliptic partial differential equations.* J. Funct. Anal. 254 (2008), no. 4, 904–933.

[47] L. Hauswirth and F. Pacard, *Higher genus Riemann minimal surfaces*, Invent. Math. 169(2007), 569–620.

[48] D. Hoffman and H. Karcher, *Complete embedded minimal surfaces of finite total curvature.* In Geometry V, Encyclopaedia Math. Sci., vol. 90, pp. 5–93, 262–272. Springer Berlin (1997)

[49] D. Hoffman and W.H. Meeks III, *A complete embedded minimal surface in* $\mathbb{R}^3$ *with genus one and three ends*, J. Diff. Geom. 21 (1985), 109–127.

[50] D. Hoffman and W.H. Meeks III, *The asymptotic behavior of properly embedded minimal surfaces of finite topology*, J. Am. Math. Soc. 4(2) (1989), 667–681

[51] D. Jerison and R. Monneau, *Towards a counter-example to a conjecture of De Giorgi in high dimensions,* Ann. Mat. Pura Appl. 183 (2004), 439–467.

[52] Kang, X., Wei, J. *On interacting bumps of semi-classical states of nonlinear Schrödinger equations*, Adv. Diff. Eqn. 5(7–9), 899–928 (2000).

[53] C.E. Kenig, F. Merle, *Global well-posedness, scattering and blow-up for the energy-critical, focusing, non-linear Schrdinger equation in the radial case.* Invent. Math. 166 (2006), no. 3, 645–675.

[54] R. V. Kohn and P. Sternberg, *Local minimizers and singular perturbations,* Proc. Roy. Soc. Edinburgh Sect. A, 11 (1989), pp. 69–84.

[55] M. Kowalczyk, *On the existence and Morse index of solutions to the Allen-Cahn equation in two dimensions*, Ann. Mat. Pura Appl. (4), 184 (2005), pp. 17–52.

[56] M.K. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$ *in* R^n, Arch. Rational Mech. Anal. 105 (1989), no. 3, 243–266.

[57] Y.Y. Li, *On a singularly perturbed elliptic equation*, Adv. Diff. Eqn. 2(6), 955–980 (1997).

[58] F. J. López, A. Ros, *On embedded complete minimal surfaces of genus zero*, J. Differential Geom. 33 (1991), no. 1, 293–300.

[59] F. Mahmoudi, A. Malchiodi, M. Montenegro, *Solutions to the nonlinear Schrdinger equation carrying momentum along a curve.* Comm. Pure Appl. Math. 62 (2009), no. 9, 1155–1264.

[60] A. Malchiodi, M. Montenegro, *Boundary concentration phenomena for a singularly perturbed elliptic problem,* Commun. Pure Appl. Math., 55 (2002), pp. 1507–1568.

[61] A. Malchiodi, M. Montenegro, *Multidimensional boundary layers for a singularly perturbed Neumann problem,* Duke Math. J., 124 (2004), pp. 105–143.

[62] A. Malchiodi, *Some new entire solutions of semilinear elliptic equations on* $\mathbb{R}^n$. Adv. Math. 221 (2009), no. 6, 1843–1909.

[63] L. Modica, *Convergence to minimal surfaces problem and global solutions of* $\Delta u = 2(u^3 - u)$. Proceedings of the International Meeting on Recent Methods in Nonlinear Analysis (Rome, 1978), pp. 223–244, Pitagora, Bologna, (1979).

[64] L. Modica, S. Mortola, *Un esempio di* Γ*-convergenza.* Boll. Unione Mat. Ital. Sez. B 14, 285–299 (1977).

[65] F. Morabito, *Index and nullity of the Gauss map of the Costa-Hoffman-Meeks surfaces*, preprint 2008.

[66] M. Musso, F. Pacard, J. Wei. *Finite-energy sign-changing solutions with dihedral symmetry for the stationary nonlinear Schrodinger equation,* Preprint 2009.

[67] S. Nayatani, *Morse index and Gauss maps of complete minimal surfaces in Euclidean* $3-$*space*, *Comm. Math. Helv.* 68(4)(1993), 511–537.

[68] S. Nayatani, *Morse index of complete minimal surfaces. In: Rassis, T.M. (ed.) The Problem of Plateau*, pp. 181–189(1992)

[69] W.-M. Ni, I. Takagi, *On the shape of least-energy solutions to a semilinear Neumann problem*, Comm. Pure Appl. Math. 44 (1991), no. 7, 819–851.

[70] W.-M. Ni, I. Takagi, *Locating the peaks of least-energy solutions to a semilinear Neumann problem*, Duke Math. J. 70 (1993), no. 2, 247–281.

[71] W.-M. Ni, J. Wei, *On the location and profile of spike-layer solutions to singularly perturbed semilinear Dirichlet problems.* Comm. Pure Appl. Math. 48 (1995), no. 7, 731–768.

[72] M. Obata, *Conformal changes of Riemannian metrics on a Euclidean sphere.* Differential geometry (in honor of Kentaro Yano), pp. 347–353. Kinokuniya, Tokyo, (1972).

[73] F. Pacard and M. Ritoré, *From the constant mean curvature hypersurfaces to the gradient theory of phase transitions,* J. Differential Geom. 64 (2003), no. 3, 356–423.

[74] J. Pérez, A. Ros, *The space of properly embedded minimal surfaces with finite total curvature.* Indiana Univ. Math. J. 45 (1996), no. 1, 177–204

[75] P. Rabinowitz, *On a class of nonlinear Schrodinger equations.* Z. Angew. Math. Phys. 43 (1992), no. 2, 270-291.

[76] O. Savin, *Regularity of flat level sets in phase transitions.* Ann. of Math. (2) 169(2009), no.1, 41–78.

[77] M. Schatzman, *On the stability of the saddle solution of Allen-Cahns equation*, Proc. Roy. Soc. Edinburgh Sect. A 125 (1995), 1241–1275.

[78] R. Schoen, *Uniqueness, symmetry, and embeddedness of minimal surfaces*, J. Differential Geom. 18 (1983), 791–809

[79] L. Simon, *Entire solutions of the minimal surface equation*, J. Differential Geometry 30 (1989), 643–688.

[80] J. Simons, *Minimal varieties in riemannian manifolds.* Ann. of Math. (2) 88(1968), 62–105.

[81] W. A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys. 55 (1977), no. 2, 149–162

[82] G. Talenti, *Best constants in Sobolev inequality,* Annali di Matematica 10 (1976), 353–372

[83] J. Wei, *On the boundary spike layer solutions to a singularly perturbed Neumann problem.* J. Differential Equations 134 (1997), no. 1, 104–133.

[84] J. Wei and S. Yan, *Infinitly many solutions for the prescribed scalar curvature problem on* S^N, preprint 2009.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

The Solvability of Differential Equations

Nils Dencker*

Abstract

It was a great surprise when Hans Lewy in 1957 presented a non-vanishing complex vector field that is not locally solvable. Actually, the vector field is the tangential Cauchy–Riemann operator on the boundary of a strictly pseudoconvex domain. Hörmander proved in 1960 that almost all linear partial differential equations are not locally solvable. This also has connections with the spectral instability of non-selfadjoint semiclassical operators.

Nirenberg and Treves formulated their well-known conjecture in 1970: that condition (Ψ) is necessary and sufficient for the local solvability of differential equations of principal type. Principal type essentially means simple characteristics, and condition (Ψ) only involves the sign changes of the imaginary part of the highest order terms along the bicharacteristics of the real part.

The Nirenberg-Treves conjecture was finally proved in 2006. We shall present the background, the main ideas of the proof and some open problems.

Mathematics Subject Classification (2010). Primary 35A01; Secondary 35S05, 47G30, 58J40.

Keywords. Solvability, pseudodifferential operators, principal type, systems of differential equations, pseudospectrum.

1. Introduction

Sixty years ago, it was generally believed that at least non-degenerate linear partial differential equations $Pu = f$ have a local solution u for any smooth f. Particularly, since Malgrange and Ehrenpreis had proved that all constant coefficient linear partial differential equations have local solutions, and by the Cauchy–Kovalevsky Theorem all analytic partial differential equations have local analytic solutions.

*Department of Mathematics, Lund University, Box 118, SE-221 00 Lund, Sweden.
E-mail: dencker@maths.lth.se.

Therefore, it came as a complete surprise when Hans Lewy [35] showed that the complex analytic vector field

$$L = \partial_{x_1} + i\partial_{x_2} + i(x_1 + ix_2)\partial_{x_3} \tag{1.1}$$

has the property that $Lu = f$ has *no* local solution u at any point in $\mathbf{R}^3$ for almost all smooth f, not even in a weak sense. Actually, it turns out that the range of L is of the first category in C^∞ and it contains the analytic functions by the Cauchy–Kovalevsky Theorem.

What is even more surprising is that L is a non-vanishing vector field with polynomial coefficients and it is naturally occuring, since it is the tangential Cauchy–Riemann operator on the boundary of the strictly pseudoconvex domain

$$\left\{ (z_1, z_2) \in \mathbf{C}^2 : \ |z_1|^2 + 2\,\mathrm{Im}\, z_2 < 0 \right\}$$

This discovery opened up a new research area: to understand and find conditions for the solvability of differential operators. Actually, in the definition of local solvability the operator does not have to be surjective.

Definition 1.1. We say that the differential operator P is locally solvable at a point x if the equation

$$Pu = f \tag{1.2}$$

has a solution u near x for almost all $f \in C^\infty$.

Here we shall also allow weak (distributional) solutions to the equation. The condition "almost all"means that f satisfies finitely many conditions, so that the range of P has finite codimension in C^∞. Recall that a differential operator is given by

$$p(x, D)u = \frac{1}{(2\pi)^n} \int e^{i\langle x,\xi\rangle} p(x, \xi)\hat{u}(\xi)\, d\xi \tag{1.3}$$

where $\hat{u}$ is the Fourier transform of $u \in C_0^\infty(\mathbf{R}^n)$, the smooth functions which are zero outside a compact set, the support of u. By this definition, ξ_j corresponds to $D_{x_j} = \frac{1}{i}\partial_{x_j}$ so a real $p(\xi)$ gives a symmetric operator. If $\xi \mapsto p(x, \xi)$ is a polynomial of degree m then $p(x, D)$ is a partial differential operator, else one can define it as a *pseudodifferential operator*. For example, when $p(x, \xi) = |\xi|$ we get $|D| = (-\Delta)^{1/2}$. The calculus gives *classical* pseudodifferential operators with the asymptotic expansion

$$p(x, \xi) = p_m(x, \xi) + p_{m-1}(x, \xi) + \dots \tag{1.4}$$

where $p_j(x, \xi)$ is homogeneous of degree j in ξ. The function $p(x, \xi)$ is called the *symbol* of the operator $p(x, D)$, and the highest order term $p_m(x, \xi)$ is called the *principal symbol*, which is invariant under changes of variables, see [23, Chapter 18]. The use of non-polynomial symbols makes it possible to *microlocalize*, i.e., localize also in the frequency variable ξ. One can then define microlocal solvability of pseudodifferential operators, see [23, Definition 26.4.3].

Now, by the Hahn–Banach and Banach Theorems, local solvability is equivalent to *a priori* estimates of the following type:

$$\|u\| = \|u\|_{(0)} \leq C\|P^*u\|_{(k-m)} + \dots \qquad u \in C_0^\infty \tag{1.5}$$

for operators of order m. Here P^* is the adjoint and we use the L^2 Sobolev norm defined by

$$\|u\|_{(s)}^2 = \frac{1}{(2\pi)^n}\int (1+|\xi|^2)^s |\hat{u}(\xi)|^2\, d\xi = \|(1+|D|^2)^{s/2}u\|^2 \tag{1.6}$$

on the Sobolev space $H_{(s)}$. We shall also allow non-local and lower order terms in (1.5), otherwise P^* would have to be injective and P surjective. In (1.5), we define $k \geq 0$ as the *loss of derivatives* for the operator. This loss means that the equation (1.2) has a solution $u \in H_{(s+m-k)}$ for almost all $f \in H_{(s)}$. If the principal symbol satisfies $p(x,\xi) \neq 0$ for $\xi \neq 0$ then the operator is elliptic and solvable with a loss of zero derivatives. When the loss satisfies $0 < k < 1$, the operator P^* is *subelliptic*, see [23, Chapter 27]. For real vector fields the loss is one, and we obtain L^2 estimates. Then, one could make the constant small in (1.5) by restricting to $u \in C_0^\infty$ with small support. For example, for $P^* = D_t$ we have

$$\|u\| \leq CT\|P^*u\| \qquad \text{when } |t| \leq T \text{ in the support of } u \tag{1.7}$$

which can be generalized to non-vanishing real vector fields by a change of variables. This makes it possible to perturb the estimate (1.5) with lower order terms in P^*.

The generic case for non-elliptic operators is when the principal symbol $p(x,\xi)$ vanishes of first order, so that the gradient satisfies $\nabla p(x,\xi) \neq 0$ when $p(x,\xi) = 0$ and $\xi \neq 0$. Then we say that the operator is of *principal type*. This terminology was first introduced by Hörmander [16] in his thesis 1955. For example, non-vanishing complex vector fields like Lewy's example (1.1) are of principal type.

Lewy's example showed that operators of principal type are not always locally solvable. This started the quest to find the conditions for the solvability of principal type operators, and it turned out that almost all non-elliptic partial differential equations are *unsolvable*. The solvability question presented an interesting and complex research field, for which many tools were developed. It actually took about 50 years to determine the precise conditions for local solvability of operators of principal type, with the final answer [9] appearing in 2006.

In this paper, we shall make a short review of the complex history of the solvability problem for differential operators of principal type and we shall only consider the smooth category. We shall outline what is known in the research area and also pose some open problems. The paper is intended as a non-technical introduction, so we shall avoid unnecessary technicalities and instead concentrate on the main ideas. For a more extensive historical review, we refer the reader to [25] (see also [32]).

2. History

Since all non-vanishing real vector fields and all constant coefficient PDE's are locally solvable, the solvability of P seems to depend on the commutator between the symmetric part $\operatorname{Re} P = (P+P^*)/2$ and antisymmetric part $\operatorname{Im} P = (P-P^*)/2i$, where P^* is the adjoint. In fact, for vector fields with non-vanishing symmetric part $\operatorname{Re} P$ we can use (1.7) with $P^* = \operatorname{Re} P$ to obtain

$$\|u\| \leq CT\|(\operatorname{Re} P)u\| \leq CT(\|Pu\| + \|P^*u\|)/2 \tag{2.1}$$

when $u \in C_0^\infty$ has support in a ball of radius T. Since we estimate with the adjoint, we shall consider the difference

$$\|Pu\|^2 - \|P^*u\|^2 = \langle [P^*, P]u, u\rangle = 2i\langle [\operatorname{Re} P, \operatorname{Im} P]u, u\rangle \tag{2.2}$$

If P has principal symbol p, the commutator $i[\operatorname{Re} P, \operatorname{Im} P]$ has principal symbol equal to the Poisson bracket

$$\{\operatorname{Re} p, \operatorname{Im} p\} = \sum_j \partial_{\xi_j} \operatorname{Re} p\, \partial_{x_j} \operatorname{Im} p - \partial_{x_j} \operatorname{Re} p\, \partial_{\xi_j} \operatorname{Im} p = H_{\operatorname{Re} p} \operatorname{Im} p \tag{2.3}$$

Here $H_{\operatorname{Re} p}$ is the Hamilton vector field of $\operatorname{Re} p$, which is the generator of the bicharacteristics of $\operatorname{Re} p$, called *semibicharacteristics* of p. Thus, non-positivity of the bracket in (2.2) would give the solvability estimate (1.5) from (2.1).

Hörmander [16] had in fact already proved in his thesis 1955 that if P is a PDO of principal type with principal symbol p satisfying

$$\{\operatorname{Re} p, \operatorname{Im} p\} \equiv 0 \tag{2.4}$$

then P is locally solvable with a loss of one derivative, showing that symmetric PDO's of principal type are locally solvable. Now condition (2.4) is not invariant under the composition of P with elliptic operators. In fact, then the principal symbol p is multiplied with an invertible factor, which adds terms containing $\operatorname{Re} p$ and $\operatorname{Im} p$ to the Poisson bracket. Thus, the value of the Poisson bracket (2.3) is only invariantly defined on the characteristics $p^{-1}(0)$.

Hörmander [17] generalized in 1960 Lewy's counterexample by proving that a necessary condition for solvability is that the principal symbol p satisfies

$$\{\operatorname{Re} p, \operatorname{Im} p\} \leq 0 \qquad \text{when } p = 0 \tag{2.5}$$

For PDO's the Poisson bracket is an odd function of ξ, so this gives the *necessary* condition for solvability of principal type PDO's, that the principal symbol p satisfies

$$\{\operatorname{Re} p, \operatorname{Im} p\} = 0 \qquad \text{when } p = 0 \tag{2.6}$$

Thus *almost all* non-elliptic PDE's are unsolvable. For example, the Lewy vector field in (1.1) has bracket equal to $2\xi_3$ which is $\neq 0$ on the characteristics. For

the operator $P = D_t + if(t, x, D_x)$, with real f homogeneous of degree one in ξ, condition (2.6) means that $\partial_t f = 0$ on $f^{-1}(0)$. Unfortunately, condition (2.6) is not sufficient for solvability, see Example 2.2.

Actually, it suffices that the Poisson bracket (2.3) has some upper bound in order to obtain solvability of the operator. In fact, making the conjugation $e^{-\lambda\phi} P e^{\lambda\phi}$ with real valued ϕ subtracts $2\lambda(H^2_{\operatorname{Re} p}\phi + H^2_{\operatorname{Im} p}\phi)$ from the bracket, giving solvability with a loss of one derivative when this is large enough. This is an example of the conjugation method: if P is solvable then APB is solvable for any invertible A and B. By choosing $A = B^{-1}$ it suffices to prove the solvability of

$$B^{-1}PB = P + R$$

Now the principal symbol of R is $H_p b/ib$, where b is the principal symbol of B. This makes it possible to change the lower order terms in the expansion of P. But the problem is that the vector field H_p will in general *not* be solvable since the bracket condition (2.6) is not satisfied in general, see (2.9). Observe that if H_p has the radial direction and b is homogeneous of degree k, then $H_p b/b$ is not dependent of b. Thus, we shall in the following assume that the Hamilton field of the principal symbol p is independent of the radial direction.

By the Banach Theorem, the range of a non-solvable operator is of the first category. Hörmander [19] proved in 1963 that the range of a non-solvable vector field determines the vector field up to right multiplication by functions. This has recently been generalized to any non-solvable principal type operator by Wittsten [43] but in general with weaker conclusions.

2.1. Principally normal operators. Obviously, there is a big gap between the sufficient condition (2.4) and the necessary condition (2.6). Hörmander [18] proved in 1960 that if P is a PDO of principal type with principal symbol p satisfying

$$\{ \operatorname{Re} p, \operatorname{Im} p \} = \operatorname{Re}(qp) \tag{2.7}$$

for some symbol q, then P is solvable with a loss of one derivative. When P is of first order, for example a vector field, we find that q in (2.7) is of order 0. Then (2.2) can essentially be estimated by

$$2 \operatorname{Re}\langle Q^* P^* u, u\rangle \le C\|P^* u\|\|u\| \tag{2.8}$$

giving solvability with a loss of one derivative by using (2.1). Observe that in this case

$$[H_{\operatorname{Re} p}, H_{\operatorname{Im} p}] = H_{\{ \operatorname{Re} p, \operatorname{Im} p \}} = \operatorname{Re} q \cdot H_{\operatorname{Re} p} - \operatorname{Im} q \cdot H_{\operatorname{Im} p} \qquad \text{on } p^{-1}(0) \tag{2.9}$$

which is the Frobenius integrability condition for the real and imaginary parts of the Hamilton vector field H_p. Thus, when $d \operatorname{Re} p$ and $d \operatorname{Im} p$ are linearly independent, the characteristics $p^{-1}(0)$ have a natural two dimensional foliation

given by H_p, called two dimensional bicharacteristics. Operators satisfying (2.7) are called *principally normal*, and as before it actually suffices that there exists $q \in C^\infty$ such that

$$\{ \operatorname{Re} p, \operatorname{Im} p \} \leq \operatorname{Re}(qp) \tag{2.10}$$

in order to estimate (2.2) and get solvability. For the operator $P = D_t + if(t, x, D_x)$ with real valued $f(t, x, \xi)$ homogeneous of degree 1 in ξ, condition (2.10) means that $\partial_t f \leq \alpha f$ with $\alpha \in C^\infty$. By Grönwall's lemma we then find that $f(t, x, \xi) < 0$ implies $f(s, x, \xi) < 0$ for $s \geq t$. In the case when $d \operatorname{Re} p$ and $d \operatorname{Im} p$ are linearly independent on $p^{-1}(0)$ we find by using the Taylor formula that the condition (2.5) is equivalent to solvability. But in the general case, the gap between the sufficient and the necessary conditions remained. Simple but instructive examples are given by the following operators.

Example 2.1. The operators $P_k = D_t - it^k|D_x|$, $(t, x) \in \mathbf{R}^2$, are solvable for any $k \in \mathbf{Z}_+$ with a loss of one derivative. The adjoint $P_k^* = D_t + it^k|D_x|$ is solvable if and only if k is even, and then with a loss of one derivative.

Here it is easy to see that condition (2.5) holds when $k > 1$, but then the operators are not principally normal. Observe that when k is even we can use the conjugation method:

$$e^{\lambda t} P_k^* e^{-\lambda t} = P_k^* + i\lambda = P_{k,\lambda}^* \tag{2.11}$$

Since $t^k|D_x| \geq 0$ we find $\operatorname{Im} P_{k,\lambda}^* \geq \lambda$ so by applying to $e^{\lambda t} u$ we obtain that

$$\lambda \|e^{\lambda t} u\|^2 \leq \operatorname{Im} \langle e^{\lambda t} P_k^* u, e^{\lambda t} u \rangle \leq \|e^{\lambda t} P_k^* u\| \|e^{\lambda t} u\|$$

By taking $\lambda > 0$ we obtain solvability of P_k with a loss of one derivative. This conjugation works for P_k with $\lambda < 0$ when k is even. In fact, by the Gårding inequality it suffices that the imaginary part of the principal symbol is semibounded. When k is odd a Fourier transform in the x variable gives that $P_k u = f \in C_0^\infty$ is equivalent to

$$\partial_t \hat{u}(t, \xi) + t^k |\xi| \hat{u}(t, \xi) = i \hat{f}(t, \xi)$$

which has the general solution

$$\hat{u}(t, \xi) = i \int_0^t e^{(s^{k+1} - t^{k+1})|\xi|/k+1} \hat{f}(s, \xi)\, ds + e^{-t^{k+1}|\xi|/k+1} \hat{u}_0(\xi) \tag{2.12}$$

Since $k+1$ is even the exponential is bounded, so the inverse Fourier transform then gives solutions to $P_k u = f$. This does not work for P_k^*, in fact, the second term in (2.12) gives solutions to $P_k u = 0$ destroying any solvability estimates of the type (1.5) for $P^* = P_k$. The operators in Example 2.1 are ΨDO's, the corresponding PDO's are the following Mizohata operators.

Example 2.2. The operator $Q_k = D_t + it^k D_x$ with $(t, x) \in \mathbf{R}^2$ and $k \in \mathbf{Z}_+$ is solvable with a loss of one derivative if (and only if) k is even.

Observe that in contrast with the Lewy vector field (1.1) the operators in Examples 2.1 and 2.2 are always solvable where $t \neq 0$, since they are principally normal there. The operators in Example 2.2 were first studied by Mizohata [37], and then by Nirenberg and Treves [38] in their study of general complex vector fields in 1963. A natural generalization is the following vector fields.

Example 2.3. The C^∞ vector field $P = D_t + ia(t,x)D_{x_1}$ with $(t,x) \in \mathbf{R} \times \mathbf{R}^n$ is solvable with a loss of one derivative if $a \geq 0$. P is principally normal if $\partial_t a = \gamma a$ with $\gamma \in C^\infty$.

Nirenberg and Treves [38] proved this in the case when $t \mapsto a(t,x)$ only has zeroes of finite order. The solvability of the operators in Examples 2.2 and 2.3 can now be proved directly from the Nirenberg–Treves Lemma (Lemma 2.6 below).

The operators in Examples 2.1–2.3 are on the form

$$P = D_t + if(t,x,D_x) \qquad (t,x) \in \mathbf{R} \times \mathbf{R}^n \tag{2.13}$$

with real $f(t,x,\xi)$ homogeneous of degree 1 in ξ. They are solvable if and only if

$$t \mapsto f(t,x,\xi) \qquad \text{does } \textit{not} \text{ change sign from } - \text{ to } + \text{ for increasing } t \tag{2.14}$$

Here, the t lines are the bicharacteristics of the symmetric part D_t of the operators, called semibicharacteristics. Of course, any reasonable condition for solvability has to be invariant under multiplication with non-vanishing scalars and changes of variables.

2.2. The Nirenberg–Treves conjecture.

Condition (2.14) was generalized by Nirenberg and Treves [39] in 1970 to the following definition.

Definition 2.4. We say that p satisfies condition (Ψ) if the imaginary part of p does not change sign from $-$ to $+$ on the oriented bicharacteristics of the real part.

The bicharacteristics are the flow-out of the Hamilton vector field of $\operatorname{Re} p$ on $(\operatorname{Re} p)^{-1}(0)$, they are called semibicharacteristics of p and have a natural orientation. For the model operators (2.13), condition (Ψ) means exactly (2.14). But since the gradient of $\operatorname{Re} p$ could vanish, one also have to check the condition on ip. Actually, Nirenberg and Treves [39] proved the non-trivial fact that condition (Ψ) is invariant under multiplication with non-vanishing factors, so it suffices to check the condition for some $z \in \mathbf{C}$ such that $d\operatorname{Re}(zp) \neq 0$. Observe that the adjoint P^* has principal symbol $\overline{p}$, which can then only have sign changes from $-$ to $+$. Condition (Ψ) has an interesting geometrical interpretation, see [42].

Example 2.5. The operator $P = D_t + if(t,x)|D_x|$, $(t,x) \in \mathbf{R} \times \mathbf{R}^n$, satisfies condition (Ψ) if and only $f(t,x)$ does not change sign from $-$ to $+$ as t increases.

For exampel, the operators P_k and P_k^* in Example 2.1 both satisfy condition (Ψ) when k is even, and P_k but not P_k^* satisfies the condition when k is odd. Since the Poisson bracket $\{\operatorname{Re} p, \operatorname{Im} p\} = H_{\operatorname{Re} p} \operatorname{Im} p$ is the derivative of $\operatorname{Im} p$ along the bicharacteristics of $\operatorname{Re} p$, condition (Ψ) is stronger than (2.5) but weaker than (2.10). Nirenberg and Treves [39] made the following famous conjecture about local solvability in 1970.

Conjecture (The Nirenberg–Treves conjecture). A pseudodifferential operator of principal type is locally solvable if and only if the principal symbol satisfies condition (Ψ).

For PDO's condition (Ψ) is equivalent to condition (P) which rules out *any* sign changes of the imaginary part on the bicharacteristics of the real part. This follows from the parity of the principal symbol: if the order m is odd the principal symbol changes sign when we switch ξ to $-\xi$, if it is even the bicharacteristics change direction. The Mizohata operator Q_k in Example 2.2 satisfies condition (P) if and only if k is even. The operator $P = D_t + if(t, x, D_x)$ with real $f(t, x, \xi)$ homogeneous of degree 1 in ξ, satisfies condition (P) if and only if $t \mapsto f(t, x, \xi)$ does not change sign. Actually, Nirenberg and Treves formulated condition (P) already in their study [38] of complex vector fields in 1963. Vector fields satisfying condition (P) can by a change of variables be put on the form in Example 2.3 with D_{x_1} replaced with a real vector field $B(x, D_x)$ which is constant in t.

Nirenberg and Treves [39] proved Conjecture 2.2 for PDO's having analytic principal symbol, getting a loss of one derivative. In the proof they used microlocal analysis and the Weierstrass Preparation Theorem to reduce to the model operator (2.13). By using analyticity and condition (P) they could then factorize the imaginary part $f(t, x, \xi)$ and reduce to the following lemma.

Lemma 2.6 (The Nirenberg–Treves Lemma). *Assume that* $P^* = D_t + iA(t)B$ *where* $0 \le A(t) \le C$, $B = B^*$ *is self-adjoint and the commutators* $[B, A(t)]$ *and* $[B, [B, A(t)]]$ *are bounded in* L^2. *Then*

$$\|u\| \le CT\|P^* u\| \tag{2.15}$$

if $u \in C_0^\infty$ *has support where* $|t| \le T \ll 1$.

When $B \ge 0$ one can conjugate as in (2.11) to obtain the estimate (2.15). In general, one gets (2.15) by applying the operator on the positive and non-positive eigenspaces of B, using the commutator conditions. As before, the estimate (2.15) can be perturbed with bounded operators for small enough T.

Actually, the analyticity of the principal symbol was only used to get the factorization needed in Lemma 2.6. But such a factorization is in general not possible in the smooth case as shown by a counterexample by Treves, see [20, p. 576].

2.3. The Beals–Fefferman localization. Beals and Fefferman [1] managed to get around the factorization problem in 1973 by an innovative localization, proving that condition (P) is sufficient for local solvability.

Theorem 2.7. *[1, Theorem 1] Assume that P is a pseudodifferential operator of principal type with principal symbol satisfying condition (P). Then P is solvable with a loss of one derivative.*

Actually, Beals and Fefferman proved this for partial differential operators, but the proof is microlocal and works for pseudodifferential operators as well. We shall give a short sketch of the main ideas of the proof.

First, by using microlocal analysis and the Malgrange Preparation Theorem, one can reduce the adjoint to the first order model operator $P^* = D_t + if(t,x,D_x)$, with real $f(t,x,\xi)$ homogeneous of degree one in ξ. By a Calderón–Zygmund decomposition one can localize where $|\xi| \approx h^{-1} \gg 1$ so that $|f| = \mathcal{O}(h^{-1})$. By homogeneity one obtains that

$$\partial_x^\alpha \partial_\xi^\beta f(t,x,\xi) = \mathcal{O}(h^{|\beta|-1}) \tag{2.16}$$

which is the usual classical symbol estimates. Here we use the multi-index notation $\partial_x^\alpha = \partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n}$ and $|\alpha| = \alpha_1 + \dots + \alpha_n$ for $\alpha \in \mathbf{N}^n$. Now, by choosing $h^{-1/2}x$ as new coordinates, one can make the symbol estimates isotropic:

$$\partial_x^\alpha \partial_\xi^\beta f(t,x,\xi) = \mathcal{O}(h^{-1+(|\alpha|+|\beta|)/2}) \tag{2.17}$$

in boxes $|\Delta x| + |\Delta\xi| \le Ch^{-1/2}$.

The revolutionary idea of Beals and Fefferman was to localize in even smaller boxes $|\Delta x| + |\Delta\xi| \le CH^{-1/2}$ so that f is still of first order:

$$\partial_x^\alpha \partial_\xi^\beta f(t,x,\xi) = \mathcal{O}(H^{-1+(|\alpha|+|\beta|)/2}) \tag{2.18}$$

Here the length scale depends on the symbol:

$$1 \le H^{-1} = |f| + |f'|^2 + 1 \le Ch^{-1} \tag{2.19}$$

which actually are the smallest neighborhoods such that this works. Observe that the calculus requires that we take the length scale $H^{-1/2} \ge 1$, which is the well-knowned uncertainty principle. For $|\alpha| + |\beta| = 0$ or 1, the estimate (2.18) follows directly from the definition (2.19), for higher order derivatives it follows from (2.17) and the fact that $h \le CH$. Beals and Fefferman made the localization constant in t by taking the supremum of H^{-1} over t. If this supremum has a fixed bound, the imaginary part is bounded and one obtains estimates by the conjugation method. If $\sup H^{-1} \approx H^{-1}(t_0) \gg 1$ then $f(t_0,x,\xi)$ is either elliptic or of principal type in a neighborhood. By using the Taylor formula and condition (P), one then obtains the factorization in the Nirenberg–Treves Lemma with $B = f(t_0,x,D_x)$.

This idea of microlocalizing in neighborhoods depending on a specific operator, was developed by Beals and Fefferman [2] into the well-knowned Beals–Fefferman calculus. It was then generalized by Hörmander [21] into the renowned *Weyl calculus* in 1979.

One can also define semiglobal solvability of differential equations. This means that the equation $Pu = f$ has a solution u near a given compact set K for almost all f. Then one also needs geometrical conditions, for example that there are no closed bicharacteristics of P in K, as shown by the angular vector field on an annulus in the plane. Hörmander [20] studied in 1978 the semiglobal solvability of pseudodifferential operators of principal type. If the operator satisfies condition (P) and the non-trapping condition that there are no closed bicharacteristics over a compact set K, Hörmander proved that operator is semiglobally solvable at K with a loss of $1 + \varepsilon$ derivative for any $\varepsilon > 0$. One also gets C^∞ solvability, giving a smooth solution u for smooth f. Since there are no simple normal forms in the semiglobal case, the proof instead studies the propagation of microlocal singularities (wave front sets) for the adjoint P^*. In fact, if $P^*u \in C^\infty$ then the *regularity function*

$$s_u^*(x,\xi) = \sup \{\, s : \ u \in H_{(s)} \text{ microlocally at } (x,\xi) \,\} \tag{2.20}$$

is essentially superharmonic on the two dimensional characteristics with respect to the complex structure given by the Hamilton vector field H_p, see (2.9). Observe that condition (P) implies that the bracket $\{ \operatorname{Re} p, \operatorname{Im} p \} = 0$ on $p^{-1}(0)$. On one dimensional characteristics, where H_p is proportional to a real vector field, s_u^* is quasiconcave so that the infimum is attained at the endpoints. Actually, the final case when the imaginary part vanishes of at least third order was proved in 1981 by Dencker [3]. Since $u \in C^\infty$ microlocally if and only if $s_u^* = \infty$ in a neighborhood, the wave front set is a union of one and two dimensional bicharacteristics. By the non-trapping condition all solutions with support in K to $P^*u = 0$ are smooth, which gives semiglobal solvability of P. But since we are taking the supremum in (2.20) one loses $\varepsilon > 0$ derivatives when using the regularity function.

In 1980 Hörmander [22] proved the necessity of condition (Ψ) for local and semiglobal solvability of principal type operator:

Theorem 2.8. *[22, Theorem 3.2] Assume that P is a pseudodifferential operator of principal type and that P is solvable at the compact set K. Then the principal symbol symbol of P satisfies condition (Ψ) near K.*

Thus, the case for PDO's is closed: condition (P) is equivalent to local solvability, and the loss is one derivative. The approach of the proof of Theorem 2.8 is to construct approximate solutions to the adjoint equation $P^*u = 0$ when condition (Ψ) does not hold, using an idea by Moyer. This involved modifying the geometrical optics method to the complex case. In general, this is complicated because the Hamilton vector field of the principal symbol is not solvable since it does not satisfy condition (P).

But the important case of pseudodifferential operators remained, as well as the microlocal solvability of partial differential operators. Lerner [26] proved that condition (Ψ) is sufficient for local solvability with a loss of one derivative in the two dimensional case (and in the oblique derivative problem [27]), by factorizing the imaginary part as in Example 2.5. Menikoff [36] had already in 1974 showed that an operator on the normal form $P = D_t + if(t, x, D_x)$ is locally solvable with a loss of one derivative if it satisfies condition (Ψ) and also

$$|\partial_x f|^2 + |\partial_\xi f|^2 \leq C|f| \qquad |\xi| = 1 \tag{2.21}$$

which follows from Glaeser's inequality if $(x, \xi) \mapsto f(t, x, \xi)$ does not change sign for fixed t as in Example 2.10 below. Thus, since condition (Ψ) is sufficient for local solvability of principal type operators with a loss of one derivative in two dimensions and for PDO's, it was believed that this was true in general, and there even appeared claims for this.

2.4. Lerner's counterexample. Therefore it came as a surprise when Lerner [28] in 1994 presented a counterexample: a first order principal type operator P in $\mathbf{R}^3$ that satisfies condition (Ψ) but is not locally solvable with a loss of one derivative near the origin. In fact, there exists $u_j \in C_0^\infty$ such that $|x| \leq 1/j$ in the support of $u_j(x)$ and

$$\lim_{j\to\infty} \|u_j\|/\|P^* u_j\| = \infty \tag{2.22}$$

The adjoint of P is on the form

$$P^* = D_t + iA(t)B(t) + R(t) \tag{2.23}$$

where $0 \leq A(t)$ is bounded, $t \mapsto B(t) = B^*(t)$ is of order one and non-decreasing in L^2, $R(t)$ is bounded. Actually, $A(t)$ vanishes of infinite order at $t = 0$, and

$$B(t) = D_{x_1} + H(t)W$$

where W is a non-negative operator in the variables $(x_1, x_2) \in \mathbf{R}^2$ and $H(t)$ is the Heaviside function. Even though $B(t)$ is non-decreasing as an operator on L^2, the positive eigenspace of $B(t)$ for $t < 0$ is very close to the negative eigenspace of $B(t)$ for $t > 0$. This makes it possible to construct approximate solutions to $P^* u = 0$ in a similar manner as in (2.12) to obtain (2.22). This unexpected property of $B(t)$ was called the *drift* of the operator in [28], and it shows that the two dimensional operators in Examples 2.1 and 2.2 did not really exhibit the full complexity of the solvability problem. Clearly, Lerner's counterexample raised serious doubts about the validity of the Nirenberg–Treves conjecture.

The problem was now that since the loss of derivatives could be more than one, lower order terms could not be handled and the estimate (1.5) could not be

localized. In fact, localizing (1.5) with a smooth cut-off function ϕ introduces the commutator term $[P^*, \phi]$ of order $m-1$, giving

$$\|[P^*, \phi]u\|_{(k-m)} \leq C\|u\|_{(k-1)}$$

in the right hand side of estimate (1.5). This term cannot be controlled by left hand side since $k > 1$. Also lower order terms could destroy the estimate, and these could not be conjugated away since the Hamilton vector field H_p is not always solvable in the condition (Ψ) case. But Dencker [6] proved that Lerner's counterexamples are solvable with a loss of two derivatives. In fact, this follows from the following generalization of the Nirenberg–Treves Lemma:

Lemma 2.9. *[7, Theorem 2.1] Assume that P^* is on the form* (2.23) *with $0 \leq A(t) \leq C$, $t \mapsto B(t) = B^*(t)$ is non-decreasing in L^2, $R(t)$ and* $\operatorname{Im}(B(t)R(t))$ *are bounded. Then we obtain*

$$\|u\|^2 + T^2\langle ABu, Bu\rangle \leq C(T^2 \operatorname{Im}\langle P^*u, Bu\rangle + T|\langle P^*u, u\rangle|) \tag{2.24}$$

for $u \in C_0^\infty$ having support where $|t| \leq T$.

Observe that

$$\operatorname{Im}(B(t)R(t)) = [B(t), R(t)]/2i \tag{2.25}$$

if $R(r) = R^*(t)$ is symmetric, and by conjugation one can often reduce to this case. For the proof of Lemma 2.9 one observes that by using (2.1) with $P^* = D_t$ it suffices to estimate the term

$$\|ABu\|^2 \leq \|A\|\|A^{1/2}Bu\|^2 = \|A\|\langle ABu, Bu\rangle \tag{2.26}$$

If $\partial_t B \geq 0$ and $\operatorname{Im}(BR)$ is bounded, this can be done by using the identity

$$2\operatorname{Im}\langle BP^*u, u\rangle = \langle \partial_t Bu, u\rangle + 2\langle ABu, Bu\rangle + 2\operatorname{Im}\langle BRu, u\rangle \tag{2.27}$$

The adjoint of Lerner's counterexample can be put on the form of Lemma 2.9 with a first order $B(t)$ so that $\operatorname{Im}(BR)$ is bounded, which gives solvability with a loss of two derivatives. In the case $A \geq c > 0$ we obtain that

$$\operatorname{Im}\langle P^*u, Bu\rangle = \operatorname{Im}\langle A^{-1/2}P^*u, A^{1/2}Bu\rangle \leq \|A^{-1/2}P^*u\|\langle ABu, Bu\rangle^{1/2} \tag{2.28}$$

We then obtain L^2 estimates for P^* from (2.24) for small enough T by the Cauchy–Schwarz inequality. Thus it is essential that $A(t)$ vanishes in Lerner's counterexample.

Lemma 2.9 gives solvability for P with a loss of two derivatives in the case when one can factorize the imaginary part so that

$$P^* = D_t + ia(t, x, D_x)b(t, x, D_x) + r(t, x, D_x) \tag{2.29}$$

where $a \geq 0$ and r are bounded symbols, b is real first order and non-decreasing in t (see [7, Corollary 2.6] or [31, Theorem 2.2]). In fact, one can then conjugate

to make r real so that $\mathrm{Im}(b(t,x,D_x)r(t,x,D_x))$ is bounded by (2.25). In the case $a > 0$ one can use (2.28) to show that the loss is one derivative. Lemma 2.9 was also used in [13] to prove hypoellipticity for operators of degenerate Egorov type, where B in (2.23) also satisfies the condition that $\partial_t^k B \neq 0$ for some k.

Lemma 2.9 is an example of the *multiplier method*: to use the formal identity

$$2\,\mathrm{Im}\,BP^* = (BP^* - PB)/i = \partial_t B + 2\,\mathrm{Re}\,BF \tag{2.30}$$

for the model operator $P^* = D_t + iF(t)$ with *symmetric* $F(t)$. Then the problem is to find lower bounds on both $\partial_t B$ and $\mathrm{Re}\,BF$. This method was used by Lerner for the proof of the sufficiency of condition (Ψ) in two dimensions, but with a bounded multiplier which was essentially the *sign* of F. Important for that proof is that the sign changes of the imaginary part of the principal symbol can only occur in the x variables. Hörmander [25] generalized this to the case when the principal symbol $f(t,x,\xi)$ of $F(t)$ satisfies condition (Ψ) and only has Lagrangean sign changes, which means that the sign of $f(t,x,\xi)$ is independent of ξ. A simple example is the following result of Egorov [15] from 1974.

Example 2.10. If $P = D_t + if(t,x,D_x)$ where f is first order satisfying $tf \leq 0$, then by using the bounded multiplier $B = t$ and the Gårding inequality we obtain from (2.30) that P is locally solvable with a loss of one derivative.

Now factorizations of the form in (2.29) of the principal symbol of the imaginary part $f(t,x,\xi)$ always exist in the condition (Ψ) case. For example $f = \mathrm{sgn}(f)|f|$ where the sign function has the property that $t \mapsto \mathrm{sgn}(f)$ is non-increasing by condition (Ψ). But the problem is to have a factorization in sufficiently good symbol classes, so that (2.25) is bounded.

Lerner [30] proved that a first order principal type operator that satisfies condition (Ψ) can be written on the normal form (2.23) but with a bounded term R that does not satisfy the conditions in Lemma 2.9, since $\mathrm{Im}(BR)$ is not bounded. Thus every such operator is a sum of a solvable operator and a bounded operator. But since the loss of derivatives is greater than one for the solvable operator, perturbing with the bounded operator could destroy the solvability estimate. It is still not known whether it is always possible in the condition (Ψ) case to reduce the adjoint P^* to the form (2.23) satisfying the conditions in Lemma 2.9, except when the dimension is two.

3. The Resolution of the Nirenberg–Treves Conjecture

In 2006 Dencker [9] finally proved the sufficiency of condition (Ψ) for local solvability of principal type operators with a loss of two derivatives, which resolved the Nirenberg–Treves conjecture.

Theorem 3.1. *[9, Theorem 1.1] Assume that P is a principal type pseudodifferential operator with principal symbol satisfying condition (Ψ). Then P is locally solvable with a loss of two derivatives.*

Instead of factorizing the imaginary part, the proof involved a direct construction of multiplier B to use in an estimate of the adjoint. This was then improved by Dencker [8] to a loss of $3/2+\varepsilon$ derivatives, for any $\varepsilon>0$. Lerner [33] improved the loss to exactly $3/2$ derivatives, by using essentially the same method of proof but with an different multiplier. Observe that there are no counterexamples giving a loss of more than $1+\varepsilon$ derivatives, for any $\varepsilon>0$.

3.1. The Proof.

The proof of the Nirenberg–Treves conjecture is long and complicated, so we shall only give the main ideas of the proof. First, by using microlocal analysis and the Malgrange Preparation Theorem, one can reduce the adjoint to the first order model operator

$$P^* = D_t + if(t,x,D_x) \tag{3.1}$$

where f is real valued and homogeneous of degree one in ξ. Condition (Ψ) means that $t\mapsto f(t,x,\xi)$ does not change sign from $+$ to $-$. Since we lose more than one derivative in the estimate, the reduction to (3.1) is rather nontrivial and only possible because of the special type of estimate that we are proving, see (3.2) below. We can also localize where $|\xi|\approx h^{-1}\gg 1$ so that $|f|=\mathcal{O}(h^{-1})$. By homogeneity one easily obtains the usual classical symbol estimates (2.16) for f. In the following, we shall avoid using explicit constants, and instead use the notation $a\lesssim b$ when $a\le Cb$. As before, by choosing $h^{-1/2}x$ as new coordinates, one can make the symbol estimates isotropic as in (2.17), in particular $\partial^3_{x,\xi}f=\mathcal{O}(h^{1/2})$. We can also localize in neighborhoods where $|\Delta x|+|\Delta\xi|\lesssim h^{-1/2}$.

Claim 3.2. Theorem 3.1 will follow if we can find a symmetric multiplier B such that $\|B\|\lesssim h^{-1/2}$ and

$$h^{1/2}\|u\|^2 \lesssim T\,\mathrm{Im}\langle P^*u, Bu\rangle \tag{3.2}$$

when $u\in C_0^\infty$ has support where $|t|\le T$.

In fact, we then obtain

$$h^{1/2}\|u\|^2 \lesssim Th^{-1/2}\|P^*u\|\|u\|$$

which gives the solvability estimate with a loss of two derivatives for the original operator. Now, since we are taking the imaginary part in (3.2), this estimate can be localized and perturbed with lower order terms. In fact, lower order terms can be made symmetric by conjugation, since $\mathrm{Re}\,P\sim D_t$ is solvable. Since B essentially is of order $1/2$ we then obtain for symmetric R of order 0 that

$$\mathrm{Im}\langle Ru, Bu\rangle = \frac{1}{2i}\langle [B,R]u,u\rangle \lesssim h^{1/2}\|u\|^2$$

Now, in order to get the estimate (3.2) we shall again use the formal identity (2.30):

$$2\operatorname{Im} BP^* = (BP^* - PB)/i = \partial_t B + 2\operatorname{Re} Bf \tag{3.3}$$

so we have to find lower bounds on both $\partial_t B$ and $\operatorname{Re} Bf$. In order to avoid technicalities, we shall treat the operators as if they were functions (which is approximately true). Thus we want that $t \mapsto B(t)$ is non-decreasing and $Bf \geq 0$, one example is the sign function $\operatorname{sgn}(f)$. Another example with a more regular symbol is:

$$\delta = \operatorname{sgn}(f) \cdot d \tag{3.4}$$

where $d(t, x, \xi)$ is the (x, ξ) distance to the sign changes of $f(t, x, \xi)$ for fixed t. If f has no sign changes within the distance $h^{-1/2}$, we put $d = h^{-1/2}$ which gives $\delta = \mathcal{O}(h^{-1/2})$. By condition (Ψ) we find that the distance to the sign changes decreases when $f < 0$ and increases when $f > 0$, thus we obtain that

$$\partial_t \delta \geq 0 \quad \text{and} \quad \delta f = d|f| \geq 0$$

But the problem with choosing $B = \delta$ is that we don't get any positive lower bound on $\operatorname{Im} BP^*$, in fact, lower order terms in P^* will give a negative lower bound. That problem can be remedied by adding a strictly increasing perturbation ϱ so that $\partial_t B \geq \partial_t \varrho > 0$, see Section 3.3 below. The main problem is that $(x, \xi) \mapsto \delta(t, x, \xi)$ is not smooth, only Lipschitz continuous. But it is a well-known fact that the signed distance function is smooth when the gradient of f is non-vanishing, and then in a neighborhood that is proportional to the inverse curvature of the characteristics $f^{-1}(0)$. We shall first localize in such neighborhoods.

3.2. The localization. One of the main ideas of the proof is to localize in smaller neighborhoods $|\Delta x| + |\Delta \xi| \lesssim H^{-1/2}$ as Beals and Fefferman did in the proof of Theorem 2.7, but now with a t dependent localization.

Definition 3.3. For fixed t we define

$$1 \leq H^{-1/2} = 1 + |\delta| + \frac{|f'|}{\|f''\| + h^{1/4}|f'|^{1/2} + h^{1/2}} \lesssim h^{-1/2} \tag{3.5}$$

Here δ is the signed distance function given by (3.4), f' and f'' are the gradient and Hessian of f with respect to (x, ξ).

Then since $\partial^3_{x,\xi} f = \mathcal{O}(h^{1/2})$ by (2.17) and $h \lesssim H$ it follows that

$$\partial_x^\alpha \partial_\xi^\beta f = \mathcal{O}(MH^{(|\alpha|+|\beta|)/2}) \tag{3.6}$$

if we define

$$M = |f| + |f'|H^{-1/2} + \|f''\|H^{-1} + h^{1/2}H^{-3/2} \lesssim h^{-1} \tag{3.7}$$

By [9, Proposition 3.8] we have $M \approx \|f''\|H^{-1} + h^{1/2}H^{-3/2} \lesssim H^{-1}$. The following geometrical property will be important for the proof.

Proposition 3.4. *[9, Proposition 3.9] If $H^{-1/2} \gg 1$ when $f = 0$ then we find $|f'| \gg h^{1/2} > 0$ and the curvature of $f^{-1}(0)$ is bounded by $CH^{1/2}$. We can then factorize*

$$f = \alpha\delta \qquad \text{when } |\delta| \lesssim H^{-1/2}$$

where $MH^{1/2} \approx \alpha \in C^\infty$ so $\delta \in C^\infty$ in (x, ξ).

Now, if we add a strictly increasing perturbation ϱ to $B = \delta$ we may destroy the non-negativity of Bf. In fact, when $f = \alpha\delta$ as in Proposition 3.4 we find

$$Bf = (\delta + \varrho)\alpha\delta = \alpha(\delta + \varrho/2)^2 - \alpha\varrho^2/4 \geq -\alpha\varrho^2/4 \tag{3.8}$$

by completing the square. Since $\alpha \approx MH^{1/2}$ and $M \lesssim H^{-1}$ we find that

$$0 \leq \alpha\varrho^2 \lesssim MH^{1/2}\varrho^2 \lesssim H^{-1/2}\varrho^2$$

A first choice for the perturbation would be

$$\varrho = \frac{1}{T}\int_0^t H^{1/2}(s)\, ds \qquad |t| \leq T$$

making

$$\partial_t B \geq \partial_t \varrho \geq H^{1/2}/T \gtrsim h^{1/2}/T$$

for some $T \ll 1$, which would then give the desired lower bound in (3.2). But since $Bf \geq -CH^{-1/2}\varrho^2$ by (3.8), this can only be compensated by $\partial_t \varrho$ if

$$\varrho = \frac{1}{T}\int_0^t H^{1/2}(s)\, ds \lesssim H^{1/2}(t) \qquad |t| \leq T \tag{3.9}$$

which in general is not true, since $t \mapsto H^{1/2}(t)$ could have a large variation.

3.3. The weight.

The main problem is to, for a chosen positive weight m, find a perturbation $\varrho = \mathcal{O}(m)$ such that $\partial_t(\delta + \varrho) \gg m$.

Proposition 3.5. *[9, Proposition 5.8] For $m > 0$ we define*

$$\varrho(t) = \sup_{-T \leq s \leq t} \left(\delta(s) - \delta(t) + \frac{1}{2T}\int_s^t m(r)\, dr - m(s) \right) \qquad |t| \leq T \ll 1 \tag{3.10}$$

Then we obtain that

$$\partial_t(\delta + \varrho) \geq \frac{m}{2T} \gg m \tag{3.11}$$

We find that $|\varrho| \leq m$ if

$$\sup_{s \leq r \leq t} m(r) \leq \delta(t) - \delta(s) + m(s) + m(t) \qquad -T \leq s \leq t \leq T \tag{3.12}$$

Condition (3.12) means that $\delta(t)$ has to increase when $m(t)$ has a large variation, which is not always true for $m = H^{1/2}$. For the proof of Proposition 3.5 we observe that

$$\delta(t) + \varrho(t) = \sup_{-T\le s\le t} \left(\delta(s) - \frac{1}{2T}\int_0^s m(r)\,dr - m(s)\right) + \frac{1}{2T}\int_0^t m(r)\,dr$$

and since the supremum is non-decreasing we obtain (3.11). We get $\varrho(t) \ge -m(t)$ by putting $s = t$ in the supremum (3.10). We obtain the upper bound $\varrho \le m$ if

$$\delta(s) - \delta(t) + \frac{1}{2T}\int_s^t m(r)\,dr - m(s) \le m(t)$$

when $-T \le s \le t \le T$, which follows from (3.12) since then

$$\frac{1}{2T}\int_s^t m(r)\,dr \le \sup_{s\le r\le t} m(r) \le \delta(t) - \delta(s) + m(s) + m(t) \tag{3.13}$$

One way of obtaining (3.12) is to make the following construction.

Definition 3.6. Let

$$m(t) = \inf_{-T\le t_-\le t\le t_+\le T} \Big(\delta(t_+) - \delta(t_-) + \max\big(H^{1/2}(t_-)\langle\delta(t_-)\rangle, H^{1/2}(t_+)\langle\delta(t_+)\rangle\big)\Big) \tag{3.14}$$

where $\langle\delta\rangle = 1 + |\delta|$ and the first term is non-negative by the monotonicity of $t \mapsto \delta$.

The term $\langle\delta\rangle$ in (3.14) does not change the weight close to the sign changes of f, but makes the estimate (3.2) easier to prove when $\langle\delta\rangle \gg 1$. (Lerner used $H^{1/2}\langle\delta\rangle^2$ instead of $H^{1/2}\langle\delta\rangle$ in the definition of m in [33], which improved the estimate (3.2).) When δ is constant in t we obtain that m is quasiconvex, i.e., the supremum of m on any interval is attained at the boundary.

Proposition 3.7. *[9, Proposition 5.7] The weight m defined by* (3.14) *satisfies* (3.12) *and*

$$h^{1/2}\langle\delta\rangle \lesssim m \lesssim H^{1/2}\langle\delta\rangle \tag{3.15}$$

To prove Proposition 3.7 we observe that since $h^{1/2} \lesssim H^{1/2}$ we obtain (3.15) from Definition 3.6 by taking $t_\pm = t$. We also obtain from the definition that

$$\inf_t \Big(|\delta(t) - \delta(t_0)| + H^{1/2}(t)\langle\delta(t)\rangle\Big) \le m(t_0) \qquad \forall\, t_0$$

Since the infima in $t_\pm$ are taken independently, we obtain for fixed $s \le r \le t$ that

$$\begin{aligned} m(r) \le \inf_{t_-\le s<t\le t_+} \Big(\delta(t_+) - \delta(t_-) + \max\big(H^{1/2}(t_-)\langle\delta(t_-)\rangle, H^{1/2}(t_+)\langle\delta(t_+)\rangle\big)\Big) \\ \le \delta(s) - \delta(t) + m(s) + m(t) \end{aligned}$$

which gives (3.12) and Proposition 3.7. By taking $B = \delta + \varrho = \mathcal{O}(H^{-1/2})$ we obtain from (3.8) and (3.15) that when $H^{-1/2} \gg 1$ and $\delta = \mathcal{O}(1)$ we have

$$Bf \geq -CH^{-1/2}m^2 \geq -C_0 m \tag{3.16}$$

Away from sign changes we get a non-negative lower bound and when $H^{-1} \leq C$ we find $|BF| \lesssim MH^{3/2} \lesssim m$ by Proposition 3.8 below.

3.4. The Wick Calculus. Now to ensure that positivity of symbols gives positivity of the corresponding operators, we shall use the Wick quantization

$$a^{Wick}(x,D)u(x) = \int_{T^*\mathbf{R}^n} a(y,\eta)\Sigma_{y,\eta}(x,D)u(x)\,dyd\eta \qquad u \in C_0^\infty$$

where $\Sigma_{y,\eta}(x,D_x)$ are rank one orthogonal projections in L^2, thus $a \geq 0$ gives $a^{Wick} \geq 0$ in L^2. Then by using $B = \delta + \varrho$ we obtain from (3.11) that

$$\partial_t B^{Wick} \geq m^{Wick}/2T \gtrsim h^{1/2}/2T \tag{3.17}$$

By choosing $\Sigma_{y,\eta}$ suitably, we obtain that $a^{Wick}(x,D) \sim a(x,D)$ modulo lower order terms when a is a symbol.

But the main difficulty is that $Bf \geq -Cm$ does not imply that

$$\operatorname{Re} B^{Wick} f(t,x,D_x) \geq -Cm^{Wick} \tag{3.18}$$

because we also have to consider the lower order terms in the expansion of the composition $B^{Wick}f$. Since $|Bf| \lesssim MH^{-1/2}$ and the symbols are real, the following result will give (3.18) and thus Claim 3.2 by (3.17) and (3.3).

Proposition 3.8. *[9, Proposition 5.5] With H defined by* (3.5)*, M by* (3.7) *and m by* (3.14) *we have that*

$$MH^{3/2} \lesssim m \tag{3.19}$$

This is in fact an essential part of the proof of Theorem 3.1. Since $M \approx \|f''\|H^{-1} + h^{1/2}H^{-3/2}$, the estimate (3.19) follows from (3.15) if we can show that

$$\|f''\|H^{1/2} \lesssim m$$

By using the definition of $H^{-1/2}$ in (3.5) we obtain this if

$$\|f''\| \lesssim m^{1/2}|f'|^{1/2} + m \qquad \text{at } (t,x,\xi) \in f^{-1}(0) \tag{3.20}$$

Since $\|f''\|$ is bounded, we only have to prove (3.20) when $|f'| \lesssim m^{-1}$ and $m \ll 1$ for a fixed t which we assume equal to 0.

Now, when $m(0) \ll 1$ there exist $t_- \leq 0 \leq t_+$ by Definition 3.6 so that $0 \leq \delta(t_+) - \delta(t_-) < m$ and $H^{1/2}(t_\pm) \leq m$. This means that in a ball of radius

$\mathcal{O}(m^{-1})$ the curvature of $f^{-1}(0)$ is $\mathcal{O}(m)$ when $t = t_\pm$. By using condition (Ψ) one can find orthogonal coordinates $w = (w_1, w')$ in the (x, ξ) variables so that $f(0) = 0$

$$\operatorname{sgn}(w_1) f(0, w) \geq 0 \qquad \text{when} \quad m \cdot (1 + |w'|^2) \lesssim |w_1| \lesssim m^{-1}$$

Then by estimating the odd and even terms in the Taylor expansion we find that

$$|\partial_{w_1}^2 f(0)| \leq |f'(0)|/|w_1| + Ch^{1/2}|w_1| \qquad m \lesssim |w_1| \lesssim m^{-1}$$

By choosing $|w_1| \approx m + |f'(0)|^{1/2} m^{-1/2} \lesssim m^{-1}$ we obtain (3.20) for $\partial_{w_1}^2 f(0)$. The other second order derivatives can be estimated similarly, and we obtain Proposition 3.8 which completes the sketch of the proof of Theorem 3.1.

Note that (3.20) essentially is Glaeser's inequality for f' and it has the geometric consequence that there cannot appear any singularities in the interior of the bicharacteristics when condition (Ψ) holds.

4. Outlook and Open Problems

4.1. Solvability of systems. Since now the conditions for local solvability of scalar principal type differential operators are known, it is natural to look to the more general solvability problem for systems. To avoid complications, we shall only consider square systems of differential operators, defined as in (1.3) with $P(x, \xi)$ having values in $N \times N$ matrices. But since the results are local, they easily carry over to operators on vector bundles.

We shall define the property of being of *principal type* for systems. In the following, we shall denote by $\operatorname{Ker} P$ the kernel, $\operatorname{Ran} P$ the range and $\operatorname{Coker} P = \mathbf{C}^N / \operatorname{Ran} P$ the cokernel of the matrix P.

Definition 4.1. The $N \times N$ system $P(x, \xi) \in C^\infty$ is of principal type if

$$\operatorname{Ker} P(x, \xi) \ni u \mapsto \partial_\nu P(x, \xi) u \in \operatorname{Coker} P(x, \xi) \qquad u \in \mathbf{C}^N \tag{4.1}$$

is bijective for some ν.

Principal type implies that the algebraic and geometric multiplicities of the eigenvalue close to zero are equal when constant, see [11, Proposition 2.10].

Example 4.2. Define the system

$$P = \begin{pmatrix} \lambda & \alpha \\ 0 & \lambda \end{pmatrix}$$

with λ and $\alpha \in C^\infty$. Then P is of principal type if and only if $d\lambda \neq 0$ and $\alpha = 0$ at $\lambda^{-1}(0)$. In fact, if $\alpha \neq 0$ at $\lambda^{-1}(0)$ then $\operatorname{Ker} P = \operatorname{Ran} P = \mathbf{C} \times \{0\}$, which is preserved by $\partial_\nu P$.

Proposition 4.3. *If $P \in C^\infty$ is of principal type and A, $B \in C^\infty$ are invertible then APB is of principal type. We have that P is of principal type if and only if the adjoint P^* is of principal type.*

In fact, by Leibniz' rule we have

$$\partial_\nu(APB) = (\partial_\nu A)PB + A(\partial_\nu P)B + AP\partial_\nu B \tag{4.2}$$

and $\operatorname{Ran}(APB) = A(\operatorname{Ran} P)$ and $\operatorname{Ker}(APB) = B^{-1}(\operatorname{Ker} P)$ when A and B are invertible, which gives the invariance. Since $\operatorname{Ker} P^* = \operatorname{Ran} P^\perp$ we find that P satisfies (4.1) if and only if

$$\operatorname{Ker} P \times \operatorname{Ker} P^* \ni (u, v) \mapsto \langle \partial_\nu P u, v\rangle \tag{4.3}$$

is a non-degenerate bilinear form. Since $\langle \partial_\nu P^* v, u\rangle = \overline{\langle \partial_\nu P u, v\rangle}$ we then obtain that P^* is of principal type.

Recall that the eigenvalues of the symbol $P(x,\xi)$ are the solutions to the characteristic equation

$$|P(x,\xi) - \lambda \operatorname{Id}_N| = 0$$

where $|A|$ is the determinant of the matrix A. Now if the matrix $P(x,\xi)$ is continuous then the eigenvalues can be chosen as continuous functions. Such a continuous function $\lambda(x,\xi)$ of eigenvalues we will call a *section of eigenvalues of* $P(x,\xi)$. If the section of eigenvalues $\lambda(x,\xi)$ has constant *algebraic* multiplicity then it is a C^∞ function by the implicit function theorem.

Definition 4.4. A square symbol $P(x,\xi) \in C^\infty$ has constant characteristics if there exists an $\varepsilon > 0$ such that any section of eigenvalues λ of P with $|\lambda| < \varepsilon$ has both constant algebraic and geometric multiplicity. We say that a square system of pseudodifferential operators has constant characteristics if the principal symbol has constant characteristics.

If P has constant characteristics then any section of eigenvalues sufficiently close to zero has constant algebraic multiplicity, thus it is a C^∞ function close to zero.

Definition 4.5. Let P be a square system of pseudodifferential operators of principal type having constant characteristics. We say that P satisfies condition (Ψ) or (P) if the eigenvalue λ close to zero of the principal symbol satisfies condition (Ψ) or (P).

These conditions are well-defined, since they only concern the section of eigenvalues close to zero which is a smooth function. Similarly, one can define the condition that the Hamilton vector field H_λ of the eigenvalue λ is independent of the radial direction when $\lambda = 0$. The following result generalizes the Nirenberg–Treves conjecture to square systems.

Theorem 4.6. *[11, Theorem 2.7] Let P be a square system of pseudodifferential operators of principal type and constant characteristics, such that the Hamilton vector field H_λ of the eigenvalue λ is independent of the radial direction when $\lambda = 0$. Then P is locally solvable if and only if condition (Ψ) is satisfied, and the loss of derivatives is at most $3/2$.*

When the multiplicity of the eigenvalues of the principal symbol is not constant the situation is much more complicated. The following example shows that in general it is not sufficient to have conditions only on the eigenvalues in order to obtain solvability, not even for symmetric systems of principal type.

Example 4.7. Let $x \in \mathbf{R}^2$ and

$$P(x,D) = \begin{pmatrix} D_{x_1} & x_1 D_{x_2} \\ x_1 D_{x_2} & -D_{x_1} \end{pmatrix}$$

This system is symmetric of principal type with principal symbol having real eigenvalues $\pm\sqrt{\xi_1^2 + x_1^2\xi_2^2}$ but

$$\frac{1}{2}\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix} P \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} = \begin{pmatrix} D_{x_1} - ix_1 D_{x_2} & 0 \\ 0 & D_{x_1} + ix_1 D_{x_2} \end{pmatrix} \tag{4.4}$$

which is not solvable. In fact, the scalar operators in (4.4) do not satisfy the bracket condition (2.6) since the eigenvalues of the principal symbol of (4.4) are $\xi_1 \pm ix_1\xi_2$.

Of course, the problem is that the eigenvalues are not invariant under multiplication with invertible systems. Instead we shall introduce the following condition.

Definition 4.8. [10, Definition 4.5] The $N \times N$ system $P \in C^\infty$ is called quasisymmetrizable if there exists an $N \times N$ symmetrizer $M \in C^\infty$ and $c > 0$ so that

$$\begin{cases} \operatorname{Re}\langle M(\partial_\nu P)u, u\rangle \geq c\|u\|^2 - C\|Pu\|^2 \\ \operatorname{Im}\langle MPu, u\rangle \geq 0 \end{cases} \qquad \forall\, u \in \mathbf{C}^N \tag{4.5}$$

Note that elliptic systems are quasisymmetrizable, since one can take $M = P^{-1}$. In the scalar case, the definition means that that there exists $m \in C^\infty$ such that $\partial_\nu \operatorname{Re}(mp) > 0$ and $\operatorname{Im}(mp) \geq 0$ near $p^{-1}(0)$.

Proposition 4.9. *[10, Propositions 4.7 and 4.10] If $P \in C^\infty$ is a quasisymmetrizable system, then P is of principal type and P^* is quasisymmetrizable. If A and $B \in C^\infty$ are invertible then BPA is quasisymmetrizable.*

For quasisymmetrizable systems we have the following local solvability result.

Theorem 4.10. *[12, Theorem 2.17] Assume that P is a square system of pseudodifferential operators with quasisymmetrizable principal symbol, then P is locally solvable with a loss of one derivative.*

By using the matrix version of Malgrange Preparation Theorem in [5], one can reduce to model operators like the following example.

Example 4.11. Consider the first order system

$$P(t,x,D_t,D_x) = M(t,x,D_x)D_t + iF(t,x,D_x) \qquad (t,x) \in \mathbf{R}\times\mathbf{R}^n \tag{4.6}$$

where $M \geq c_0 > 0$ and $\operatorname{Re} F \geq 0$. Then P is quasisymmetrizable (with symmetrizer Id_N), so it is locally solvable with a loss of one derivative by Theorem 4.10.

4.2. Spectral instability. The non-solvability of differential operators has connections with the instability of spectrum (or pseudospectrum) of semiclassical operators. The spectral instability of non-selfadjoint operators is a topic of current interest in applied mathematics; it has been a problem for many years in, for example, computational fluid dynamics, see [41]. The spectral instability of semiclassical operators of principal type was studied in [14], where the important connection with the bracket condition (2.6) was made.

Definition 4.12. We define the semiclassical operator

$$p(x,hD)u(x) = \frac{1}{(2\pi)^n}\iint_{T^*\mathbf{R}^n} p\,(x,h\xi)\,e^{i\langle x-y,\xi\rangle}u(y)dyd\xi \tag{4.7}$$

When $p(x,\xi)$ is a polynomial in ξ we obtain a partial differential operator, else it could be defined as a pseudodifferential operator. One can often reduce to the case of bounded symbols $p(x,\xi) \in C^\infty_{\mathrm{b}}$, so that all derivatives $\partial_x^\alpha\partial_\xi^\beta p \in L^\infty$. This has the advantage that $p(x,hD)$ is bounded on L^2. The calculus gives an asymptotic expansion

$$P(h) = p_0(x,hD) + hp_1(x,hD) + \dots \tag{4.8}$$

where p_0 is the principal symbol of $P(h)$.

The *spectrum* of $P(h)$ is defined as the complement of the set of $z \in \mathbf{C}$ such that the resolvent $(P(h)-z)^{-1}$ exists and is a bounded operator. One can show that the spectrum of the semiclassical operator $P(h)$ in (4.8) is, for small enough h, contained in the closure of the values of the principal symbol

$$\Sigma(p) = \overline{\{\, z\in\mathbf{C} : \exists\ (x,\xi) \text{ such that } z = p_0(x,\xi)\,\}}$$

We also define the values at infinity:

$$\Sigma_\infty(p) = \{\, z\in\mathbf{C} : \exists\ (x_j,\xi_j)\to\infty,\ p(x_j,\xi_j)\to z\,\}$$

which is a compact set since p is bounded. The following definition is from [14].

Definition 4.13. We define the semiclassical pseudospectrum:

$$\Lambda(p) = \overline{\{\, p(x,\xi) :\ \{\operatorname{Re}p, \operatorname{Im}p\}\,(x,\xi)\neq 0\,\}} \subseteq \Sigma(p)$$

Observe that for analytic symbols we have that $\Lambda(p)$ is equal to either $\Sigma(p)$ or $\emptyset$. The following result shows that the spectrum is generically unstable in the values of the principal symbol when the bracket is non-zero.

Theorem 4.14. *[14, Theorem 1.2] Assume that $P(h)$ has principal symbol $p \in C^\infty_{\rm b}$. Then there exists a dense subset of $z \in \Lambda(p) \setminus \Sigma_\infty(p)$ such that*

$$\big\|(P(h) - z)^{-1}\big\| \geq C_N h^{-N} \quad \forall\, N \qquad h \to 0 \tag{4.9}$$

Here we define $\big\|(P(h) - z)^{-1}\big\| = \infty$ in the spectrum of $P(h)$. If p is analytic then h^{-N} can be replaced by $\exp(c/h)$, $c > 0$. Observe that it may happen that $\Sigma_\infty(p) = \Sigma(p)$, for example if p is constant in some variables. Actually, it suffices that condition (Ψ) is not satisfied for $p_0 - z$ in order to get (4.9). This follows by adapting the proof of the necessity of condition (Ψ) in [22], see [40]. One application is the following result for the Schrödinger equation.

Example 4.15. Let $P(h) = -h^2\Delta + V(x)$ with $V \in C^\infty(\mathbf{R}^n)$. Then, for any $z \in \{\xi^2 + V(x) : \ \mathrm{Im}\langle \xi, V'(x)\rangle \neq 0\}$ there exists $u(h) \in L^2(\mathbf{R}^n)$ with the property that $\|u(h)\| \equiv 1$ and

$$\|(P(h) - z)u(h)\| \leq C_N h^N \quad \forall\, N \qquad h \to 0 \tag{4.10}$$

which gives (4.9). If the potential $V(x)$ is real analytic then we can replace h^∞ by $\exp(-1/Ch)$ in (4.10). These approximate eigenfunctions are called *pseudomodes*.

Theorem 4.14 can be generalized to semiclassical systems of principal type, see [10]. In fact, the eigenvalues are generically of constant multiplicity, then they are C^∞ sections, so the bracket condition is well defined. We obtain as in the scalar case that the resolvent blows up as in (4.9) when the bracket is non-zero for almost all eigenvalues that are not limit eigenvalues at infinity, see [10, Theorem 3.10].

4.3. Non-linear equations.

The solvability of linear differential equations is connected with the solvability of the Cauchy problem for non-linear differential equations. In fact, the initial data determines the coefficients of the linearized equations initially, and generically the bracket condition (2.6) will not be satisfied.

Lerner, Morimoto and Xu [34] studied the instability of the C^∞ Cauchy problem for quasilinear analytic vector fields, for example Burger's equation. For any analytic initial data, the Cauchy problem for those vector fields has a local solution by the Cauchy–Kovalevsky Theorem. But for *almost all* analytic data there exists smooth data with the same Taylor expansion at a given point for which the Cauchy problem has no C^2 solution. For example, the non-homogeneous Burger's equation

$$\partial_t u + u\partial_x u = f(t, x, u) \qquad (t, x) \in \mathbf{R} \times \mathbf{R}$$

with analytic f has no C^2 solution for almost all non-analytic Cauchy data $u(0, x)$.

4.4. Open Problems. Even after the resolution of the Nirenberg–Treves conjecture, there still remain many open questions, for example, the maximal loss of derivatives in local solvability. We know that the loss is at most 3/2 derivatives, but we have no counterexample giving a loss of more than $1 + \varepsilon$ derivatives, $\forall\ \varepsilon > 0$.

Question 4.16. What is the maximal loss of derivatives for the local solvability of pseudodifferential operators of principal type satisfying condition (Ψ)?

In the condition (P) case, we obtain semiglobal solvability with a loss of arbitrarily more than one derivative. But there is no counterexample giving a loss of more than one derivative.

Question 4.17. What is the maximal loss of derivatives for the semiglobal solvability of pseudodifferential operators of principal type satisfying condition (P)?

It is not known if condition (Ψ) is sufficient for semiglobal solvability, but it is necessary by Theorem 2.8. A connected problem is the propagation of singularities for solutions to pseudodifferential equations satisfying condition (Ψ), for which little is known in general.

Question 4.18. Is condition (Ψ) is sufficient for semiglobal solvability of pseudodifferential operators of principal type?

It is also unclear what the conditions are for local solvability of pseudodifferential operators with a loss of one derivative. Condition (Ψ) is not sufficient by Lerner's counterexamples, and condition (P) is too strong. In fact, Lerner [29] proved that $P = D_t + if(t, x, D_x)$ with real first order f is solvable with a loss of one derivative if P satisfies condition (Ψ) and

$$\partial_t f \gtrsim |\partial_x f|^2 + |\partial_\xi f|^2 \qquad |\xi| = 1$$

This means that transversal sign changes give no problems. It is not known if condition (Ψ) gives a loss of one derivative in the special case when the operator has *analytic* principal symbol. Observe that since Lerner's counterexamples do not have analytic principal symbols, there are no counterexamples to this.

Question 4.19. Which are the conditions for local solvability with a loss of one derivative of pseudodifferential operators of principal type?

Very little is known about the solvability of systems of principal type having non-constant characteristics. A special case is when the principal symbol of the system is C^∞ diagonalizable, i.e., there exists a C^∞ base of eigenvectors. If then all the eigenvalues satisfy condition (P), the scalar estimates gives solvability with a loss of one derivative, since these estimates can be perturbed with any lower order terms. But when the eigenvalues have variable multiplicity satisfying condition (Ψ), one loses 3/2 derivatives in the scalar estimates, making it

impossible to perturb with any lower order term. When one can factorize the imaginary symbol as in Lemma 2.9 one gets solvability with a loss of two derivatives if lower order terms are such that (2.25) is bounded, which in general is not the case.

Question 4.20. Which are the conditions for local solvability of square systems of principal type?

Not much is known about the propagation of singularities for systems of pseudodifferential operators of principal type. For that one could use the vector valued *polarization sets* defined in [4], where the propagation of polarization for systems of real principal type was studied.

References

[1] R. Beals and C. Fefferman, *On local solvability of linear partial differential equations*, Ann. of Math. **97** (1973), 482–498.

[2] ———, *Spatially inhomogeneous pseudodifferential operators. I.* Comm. Pure Appl. Math. **27** (1974), 1–24.

[3] N. Dencker, *On the propagation of singularities for pseudodifferential operators of principal type*, Ark. Mat. **20** (1982), 23–60.

[4] ———, *On the propagation of polarization sets for systems of real principal type*, J. Funct. Anal. **46** (1982), 351–372.

[5] ———, *Preparation theorems for matrix valued functions*, Ann. Inst. Fourier (Grenoble) **43** (1993), 865–892.

[6] ———, *The solvability of non L^2 solvable operators*, Journées "Equations aux Dérivées Partielles" (Saint-Jean-de-Monts, 1996), Ecole Polytech., Palaiseau, 1996.

[7] ———, *Estimates and solvability*, Ark. Mat. **37** (1999), 221–243.

[8] ———, *The solvability of pseudo-differential operators*, Phase space analysis of partial differential equations. Vol. I, Pubbl. Cent. Ric. Mat. Ennio Giorgi, Scuola Norm. Sup., Pisa, 2004, 175–200.

[9] ———, *The resolution of the Nirenberg-Treves conjecture*, Ann. of Math. **163** (2006), 405–444.

[10] ———, *The pseudospectrum of systems of semiclassical operators*, Anal. PDE **1** (2008), 323–373.

[11] ———, *On the solvability of systems of pseudodifferential operators*, arXiv:0801.4043 [math.AP]. To appear in *Geometric Aspects of Analysis and Mechanics, A Conference in Honor of Hans Duistermaat*.

[12] ———, *The solvability and subellipticity of systems of pseudodifferential operators*, Advances in Phase Space Analysis of Partial Differential Equations, In Honor of Ferruccio Colombini's 60th Birthday, Progress in Nonlinear Differential Equations and Their Applications, Vol. 78 (A. Bove, D. Del Santo, M.K.V. Murthy, eds.), Birkhäuser, Boston, 2009, 73–94.

[13] N. Dencker, Y. Morimoto and T. Morioka, *Hypoellipticity for operators of infinitely degenerate Egorov type*, Tsukuba J. Math. **23** (1999), 215–224.

[14] N. Dencker, J. Sjöstrand, and M. Zworski, *Pseudospectra of semiclassical (pseudo-) differential operators*, Comm. Pure Appl. Math. **57** (2004), 384–415.

[15] Ju.V. Egorov, *Sufficient conditions for the local solvability of pseudodifferential equations of principal type*, Trudy Moskov. Mat. Obšč. **31** (1974), 59–83.

[16] L. Hörmander, *On the theory of general partial differential operators*, Acta Math. **94** (1955).

[17] ———, *Differential equations without solutions*, Math. Ann. **140** (1960), 169–173.

[18] ———, *Differential operators of principal type*, Math. Ann. **140** (1960), 124–146.

[19] ———, *Linear partial differential operators*, Die Grundlehren der mathematischen Wissenschaften, Bd. 116, Springer Verlag, Berlin, 1963.

[20] ———, *Propagation of singularities and semiglobal existence theorems for (pseudo)differential operators of principal type*, Ann. of Math. **108** (1978), 569–609.

[21] ———, *The Weyl calculus of pseudodifferential operators*, Comm. Pure Appl. Math. **32** (1979), 360–444.

[22] ———, *Pseudodifferential operators of principal type*, Singularities in boundary value problems (Proc. NATO Adv. Study Inst., Maratea, 1980), NATO Adv. Study Inst. Ser. C: Math. Phys. Sci., vol. 65, Reidel, Dordrecht, 1981, 69–96.

[23] ———, *The analysis of linear partial differential operators*, vol. I–IV, Springer Verlag, Berlin, 1983–1985.

[24] ———, *Notions of convexity*, Birkhäuser, Boston, 1994.

[25] ———, *On the solvability of pseudodifferential equations*, Structure of solutions of differential equations (M. Morimoto and T. Kawai, eds.), World Sci. Publ., River Edge, NJ, 1996, 183–213.

[26] N. Lerner, *Sufficiency of condition* (ψ) *for local solvability in two dimensions*, Ann. of Math. **128** (1988), 243–258.

[27] ———, *An iff solvability condition for the oblique derivative problem*, Séminaire sur les Equations aux Dérivées Partielles, 1990–1991, Ecole Polytech., Palaiseau, 1991.

[28] ———, *Nonsolvability in* L^2 *for a first order operator satisfying condition* (ψ), Ann. of Math. **139** (1994), 363–393.

[29] ———, *Energy methods via coherent states and advanced pseudo-differential calculus*, Multidimensional complex analysis and partial differential equations (São Carlos, 1995) (P. D. Cordaro, H. Jacobowitz, and S. Gidikin, eds.), Contemp. Math., vol. 205, Amer. Math. Soc., Providence, RI, 1997, 177–201.

[30] ———, *Perturbation and energy estimates*, Ann. Sci. Ecole Norm. Sup. **31** (1998), 843–886.

[31] ———, *When is a pseudo-differential equation solvable?*, Ann. Inst. Fourier (Grenoble), **50**, 2000, 443–460.

[32] ———, *Solving pseudo-differential equations*, Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002) (Beijing), Higher Ed. Press, 2002, 711–720.

[33] ———, *Cutting the loss of derivatives for solvability under condition* (Ψ), Bull. Soc. Math. France **134** (2006), 559–631.

[34] N. Lerner, Y. Morimoto and C.-J. Xu, *Instability of the Cauchy-Kovalevskaya solution for a class of non-linear systems*, Amer. J. Math. **132** (2010), 99–123.

[35] H. Lewy, *An example of a smooth linear partial differential equation without solution*, Ann. of Math. **66** (1957), 155–158.

[36] A. Menikoff, *On local solvability of pseudo-differential equations*, Proc. Amer. Math. Soc. **43** (1974), 149–154.

[37] S. Mizohata, *Solutions nulles et solutions non analytiques*, J. Math. Kyoto Univ. **1** (1961/1962), 271–302.

[38] L. Nirenberg and F. Treves, *Solvability of a first order linear partial differential equation*, Comm. Pure Appl. Math. **16** (1963), 331–351.

[39] ———, *On local solvability of linear partial differential equations. Part I: Necessary conditions*, Comm. Partial Differential Equations **23** (1970), 1–38, *Part II: Sufficient conditions*, Comm. Pure Appl. Math. **23** (1970), 459–509; *Correction*, Comm. Pure Appl. Math. **24** (1971), 279–288.

[40] K. Pravda-Starov, *Etude du pseudo-spectre d'opérateurs non auto-adjoints*, Ph.D. thesis, Université de Rennes I, 2006.

[41] L.N. Trefethen and M. Embree, *Spectra and Pseudospectra*, Princeton University Press, Princeton, N.J., 2005.

[42] F. Treves, *Winding numbers and the solvability condition* (Ψ), J. Differential Geometry **10** (1975), 135–149.

[43] J. Wittsten, *On some microlocal properties of the range of a pseudo-differential operator of principal type*, arXiv:1003.1676 [math.AP].

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Equilibrium Configurations of Epitaxially Strained Elastic Films: Existence, Regularity, and Qualitative Properties of Solutions

Nicola Fusco* and Massimiliano Morini†

Abstract

We consider a variational model introduced in the physical literature to describe the epitaxial growth of an elastic film over a thick flat substrate when a lattice mismatch between the two materials is present. We prove existence of minimizing configurations, study their regularity properties, and establish several quantitative and qualitative properties of local and global minimizers of the free-energy functional. Among the other results, we determine analytically the critical threshold for the local minimality of the flat configuration, we investigate also its global minimality, and we provide some conditions under which the non occurrence of singularities in non flat global minimizers is guaranteed. One of the main tools is a new second order sufficient condition for local minimality, which provides the first extension of the classical criteria based on the positivity of second variation to the context of functionals with bulk and surface energies.

Mathematics Subject Classification (2010). 74G55; 49K10.

Keywords. Epitaxially strained elastic films, shape instabilities, free boundary problems, second order minimality conditions, regularity

1. Introduction

The understanding of the mechanisms governing stress driven surface diffusion of atoms, which are located at the interface between two different elastic solid phases, is very important in view of its impact on several branches of physics

*N. Fusco: Dipartimento di Matematica e Applicazioni 'R. Caccioppoli', Università degli Studi di Napoli 'Federico II', Napoli, Italy. E-mail: n.fusco@unina.it.
†M. Morini: SISSA, Trieste, Italy. E-mail: morini@sissa.it.

and technology, such as low temperature physics, fracture mechanics, epitaxial growth of films, and the construction of semiconductor devices with special optic and electronic properties.

A technology that is central to the fabrication of modern semiconductor electronic and optoelectronic devices is the epitaxial deposition of a thin film onto a substrate in cases involving a mismatch in the lattice parameters between the two materials. Because of the constraint of epitaxy, a mismatch strain arises in the film and produces an interesting threshold effect: to release some of the elastic energy due to the mismatch strain, the atoms on the free surface of the film tend to diffuse and rearrange into a more favorable shape; in turn, this migration of atoms bears an energetic cost in terms of surface tension. The resulting configuration is overall more convenient only if the thickness of the film is large enough. In this case, the film profile ceases to be flat and, typically, becomes wavy or breaks into several material clusters (the so-called islands) separated by a thin layer that wets the substrate. This phenomenon is usually referred to as the Asaro-Grinfeld-Tiller instability of the flat configuration, after the name of the scientists who pioneered the theoretical investigations on this kind of stress driven morphological instabilities ([1], [11]).

Although several numerical and theoretical studies have been carried out to study qualitative properties of equilibrium configurations of strained epitaxial films (see, e.g., [10], [16], [17]), very few analytical results were present in the literature until very recently.

Perhaps, the first paper to cast the study of AGT instability in a more analytical perspective is the one by Grinfeld [12]. In the spirit of Gibbs variational approach, the author introduces a suitable free-energy functional (the sum of the stored elastic energy of the film and the interfacial energy of its free surface) and establishes various instability results for the flat morphology of the film by looking at the second variation of the free-energy. However, existence of minimizers and the problem of deriving second order sufficient minimality conditions are not addressed in that work.

Subsequently, in [3] the authors attempted to provide a well-posed variational formulation for the existence problem of minimizing configurations, but for an unrealistic one-dimensional model. Finally, in the framework of linear elasticity and considering only two-dimensional morphologies (that correspond to three-dimensional configurations with planar symmetry), the papers [2] and [8] (for a slightly different model) succeeded in determining the proper functional setting for the physically more relevant energy considered in [12]. The methods used in the two papers are related to relaxation and geometric measure theory techniques. See also [6] for a partial extension of these relaxation results to higher dimensions.

Besides dealing with the existence problem, the work [8] also develops a complete regularity theory of (locally) minimizing configurations and establishes various results about their structure and qualitative behavior. More recently, a finer study of several quantitative and qualitative properties of local

and global minimizers has been undertaken in [9]. The results of [8] and [9] will be recalled in Sections 2 and 3 of the present paper, respectively.

We now describe more in details the model studied in [2] and considered in this paper. We start by introducing the reference configuration of the film

$$\Omega_h := \left\{ z = (x, y) \in \mathbb{R}^2 : \, 0 < x < b, \, 0 < y < h\,(x) \right\} \,,$$

where $h : [0, b] \to [0, \infty)$ represents the free-profile of the film. Since we work within the theory of small deformations, we need to consider the linearized elastic strain given by

$$E\,(u) := \frac{1}{2}\left(\nabla u + \nabla^T u\right) \,,$$

with $u : \Omega_h \to \mathbb{R}^2$ representing the planar displacement.

We consider the case of a film growing on an infinitely rigid substrate. Thus, the lattice mismatch between film and substrate can be modeled by enforcing the Dirichlet boundary condition $u(x, 0) = e_0(x, 0)$ at their interface, where the positive constant e_0 is proportional to the gap between the atomic distances in the lattices of the two materials. Note that this boundary condition forces the film to be strained, thus generating elastic energy. Following the physical literature and [2], we also impose the periodicity conditions $h(0) = h(b)$ and $u(b, y) = u(0, y) + (e_0 b, 0)$. The energy associated with a configuration (h, u) when h is smooth is then given by

$$F(h, u) = \int_{\Omega_h} \left[\mu |E(u)|^2 + \frac{\lambda}{2} (\mathrm{div} u)^2 \right] \, dz + \sigma \mathcal{H}^1(\Gamma_h) \,,$$

where μ and λ represent the Lamé coefficients of the material, σ is the surface tension on the profile of the film, Γ_h is the graph of h, and $\mathcal{H}^1$ denotes the one-dimensional Hausdorff measure.

In the following, we assume without loss of generality that $\sigma = 1$. In order to describe stable equilibrium configurations, one wants to minimize F among all admissible configurations (h, u) satisfying a volume constraint $|\Omega_h| = d$. However, note that smooth sequences may converge to irregular configurations, where the profile h is just a lower semicontinuous function of bounded variation and the (extended) graph of h may contain vertical parts and cuts (the latter can be interpreted as vertical cracks forming in the film). Hence, we need to consider a larger class X of relaxed configurations and extend accordingly the definition of F, through a relaxation procedure. This has been done in [2], where it is shown that the energy associated to any pair $(h, u) \in X$ takes the form

$$F(h, u) = \int_{\Omega_h} \left[\mu |E(u)|^2 + \frac{\lambda}{2} (\mathrm{div} u)^2 \right] \, dz + \mathcal{H}^1(\Gamma_h) + 2\mathcal{H}^1(\Sigma_h) \,,$$

with Γ_h and Σ_h denoting the (extended) graph of h and the union of all vertical cuts, respectively. Notice that in the previous formula the length of vertical cuts is counted twice, since they arise as limit of regular profiles.

The plan of the paper is the following. In Section 2, after giving the precise statement of the Bonnetier-Chambolle relaxation result, we apply the regularity results established in [8] (see also [7] dealing with the case of anisotropic surface energy) to the model described above. In particular, we show that the profile of locally minimizing configurations is regular away from a finite (possibly empty) set of singularities of cusp type. As a corollary, we obtain a rigorous proof of the zero contact angle condition between film and substrate and an analytical confirmation of the formal analysis of [15]. These regularity results are in agreement with numerical simulations and experiments, where the appearance of cusps, possibly leading to vertical fractures in the material, is observed (see [10] and [16]).

In Section 3, we investigate further quantitative and qualitative properties of stable equilibrium configurations. Using the results of [9], we determine analytically the critical threshold for the local minimality of the flat configuration, we investigate also its global minimality, and we establish some conditions under which cusp singularities or fractures do not form, once the flat configuration becomes unstable. One of the main tools is a new sufficient condition for local minimality based on the positive definiteness of a suitable notion of second variation of the energy F. To the best of our knowledge, this result provides the first extension of the classical sufficiency theorems for strong local minimizers to the context of functionals with bulk and surface energies.

2. Existence and Regularity of Equilibrium Configurations

In this section we present the model studied by Bonnetier and Chambolle in [2] and the related functional setting. We also discuss the regularity theorem proved in [8].

We start by introducing the admissible profiles over the interval $(0,b)$, which are functions with finite total variation in $(0,b)$ whose b-periodic extensions are lower semicontinuous (l.s.c.) in $\mathbb{R}$. It is convenient to identify these functions with the corresponding periodic extensions:

$$AP(0,b) := \left\{ g : \mathbb{R} \to [0,+\infty) : \ g \text{ is l.s.c. and } b\text{-periodic, } \mathrm{Var}(g;0,b) < +\infty \right\},$$

where $\mathrm{Var}(g;0,b)$ denotes the *pointwise total variation* of g over the interval $(0,b)$. Since $g \in AP(0,b)$ is b-periodic, its pointwise total variation is finite over any bounded interval of $\mathbb{R}$. Therefore, it admits right and left limits at every $x \in \mathbb{R}$ denoted by $g(x+)$ and $g(x-)$, respectively. In the following we use the notation

$$g^+(x) := \max\{g(x+), g(x-)\}, \qquad g^-(x) := \min\{g(x+), g(x-)\}. \tag{1}$$

To represent the region occupied by the film, we set

$$\Omega_g := \{(x,y) : x \in (0,b),\ 0 < y < g(x)\}, \quad \Omega_g^{\#} := \{(x,y) : x \in \mathbb{R},\ 0 < y < g(x)\},$$

while the profile of the film is given by

$$\Gamma_g := \{(x,y) : x \in [0,b),\, g^-(x) \le y \le g^+(x)\}\,.$$

The set of vertical cracks (or cuts) is

$$\Sigma_g := \{(x,y) : x \in [0,b)\,,\, g(x) < g^-(x),\, g(x) \le y \le g^-(x)\}\,.$$

We will also use the notation

$$\Gamma_g^\# := \{(x,y) :\, x \in \mathbb{R},\, g^-(x) \le y \le g^+(x)\}\,.$$

The set $\Sigma_g^\#$ is defined similarly.

We now introduce a convergence in $AP(0,b)$. We recall that if A, B are closed subsets of $\mathbb{R}^2$ their Hausdorff distance is defined as

$$d_H(A,B) := \inf\{\varepsilon > 0 :\, B \subset \mathcal{N}_\varepsilon(A) \text{ and } A \subset \mathcal{N}_\varepsilon(B)\}\,,$$

where $\mathcal{N}_\varepsilon(A)$ denotes the ε-neighborhood of A.

We say that $h_n \to h$ in $AP(0,b)$ if

$$\sup_n \mathrm{Var}(h_n;0,b) < +\infty \qquad \textit{and} \qquad d_H(\mathbb{R}^2_+ \setminus \Omega^\#_{h_n}, \mathbb{R}^2_+ \setminus \Omega^\#_h) \to 0\,, \tag{2}$$

where $\mathbb{R}^2_+ = \{(x,y) \in \mathbb{R}^2 :\, y \ge 0\}$.

Given $g \in AP(0,b)$, we denote

$$LD_\#(\Omega_g;\mathbb{R}^2) := \big\{v \in L^2_{\mathrm{loc}}(\Omega_g^\#;\mathbb{R}^2) :\, v(x,y) = v(x+b,y) \text{ for } (x,y) \in \Omega_g^\#,\ E(v)|_{\Omega_g} \in L^2(\Omega_g;\mathbb{R}^2)\big\}\,,$$

where $E(v) := \frac{1}{2}(\nabla v + \nabla^T v)$, ∇v being the distributional gradient of v and $\nabla^T v$ its transpose. Given $e_0 \ge 0$, we define

$$X(e_0;b) := \big\{(g,v) :\, g \in AP(0,b),\, v : \Omega_g^\# \to \mathbb{R}^2 \text{ such that } v(\cdot,\cdot) - e_0(\cdot,0) \in LD_\#(\Omega_g;\mathbb{R}^2),\, v(x,0) = (e_0 x, 0) \text{ for all } x \in \mathbb{R}\big\}\,.$$

We introduce the following convergence in $X(e_0;b)$.

We say that $(h_n,u_n) \to (h,u)$ in $X(e_0;b)$ if and only if $h_n \to h$ in $AP(0,b)$ and $u_n \rightharpoonup u$ weakly in $H^1_{\mathrm{loc}}(\Omega_h^\#;\mathbb{R}^2)$.

Notice that the definition is well posed since by the second equation in (2) it follows that if $\Omega' \subset\subset \Omega_h^\#$ then $\Omega' \subset\subset \Omega_{h_n}^\#$ for n large enough.

We work in the framework of linearized elasticity and for simplicity we consider isotropic and homogeneous materials. Hence, the elastic energy density $Q : \mathbb{M}^{2\times 2}_{\mathrm{sym}} \to [0,+\infty)$ takes the form

$$Q(\xi) := \frac{1}{2}\mathbb{C}\xi : \xi = \mu|\xi|^2 + \frac{\lambda}{2}\big[\mathrm{tr}(\xi)\big]^2\,,$$

where

$$\mathbb{C}\xi = \begin{pmatrix} (2\mu+\lambda)\xi_{11}+\lambda\xi_{22} & 2\mu\xi_{12} \\ 2\mu\xi_{12} & (2\mu+\lambda)\xi_{22}+\lambda\xi_{11} \end{pmatrix}$$

and the *Lamé coefficients* μ and λ satisfy the ellipticity conditions

$$\mu > 0 \qquad \text{and} \qquad \lambda > -\mu\,. \tag{3}$$

If $(g,v) \in X(e_0;b)$ and g is Lipschitz the energy is defined as

$$G(g,v) := \int_{\Omega_g} Q(E(v))\,dz + \mathcal{H}^1(\Gamma_g)\,.$$

The following result, proved in [2] (see also [8]), gives a representation formula for the energy in the general case and an existence result for the corresponding constrained minimum problem. To this aim, we set for any $(g,v) \in X(e_0;b)$

$$F(g,v) := \inf\{\liminf_n G(g_n,v_n) : (g_n,v_n) \to (g,v) \text{ in } X(e_0;b),\ g_n \text{ Lipschitz}, \\ |\Omega_{g_n}| = |\Omega_g|\}\,.$$

Theorem 1. *For any pair $(g,v) \in X(e_0;b)$*

$$F(g,v) = \int_{\Omega_g} Q(E(v))\,dz + \mathcal{H}^1(\Gamma_g) + 2\mathcal{H}^1(\Sigma_g)\,. \tag{4}$$

Moreover, for any $d>0$ the minimum problem

$$\min\{F(g,v) : (g,v) \in X(e_0;b),\ |\Omega_g| = d\} \tag{5}$$

has a solution, the minimum value in (5) *is equal to*

$$\inf\{G(g,v) : (g,v) \in X(e_0;b),\ |\Omega_g| = d,\ g \textit{ Lipschitz}\}$$

and the limit points of minimizing sequences are minimizers of (5).

Our regularity result applies not only to *b-periodic global minimizers*, i.e. minimizers of (5), but also to local minimizers, which are defined as follows.

Definition 2. *We say that an admissible pair $(h,u) \in X(e_0;b)$ is a* b-periodic local minimizer *for F if there exists $\delta > 0$ such that*

$$F(h,u) \le F(g,v) \tag{6}$$

for all pairs $(g,v) \in X(e_0;b)$, with $|\Omega_g| = |\Omega_h|$ and $d_H(\Gamma_h \cup \Sigma_h, \Gamma_g \cup \Sigma_g) < \delta$. If, in addition, when $g \neq h$ (6) *holds with strict inequality, then we say that (h,u) is an* isolated b-periodic local minimizer.

Notice that a (sufficiently regular) b-periodic local minimizer $(h,u) \in X(e_0;b)$ satisfies the following set of Euler-Lagrange conditions:

$$\begin{cases} \operatorname{div} \mathbb{C}E(u) = 0 & \text{in } \Omega_h; \\ \mathbb{C}E(u)[\nu] = 0 & \text{on } \Gamma_h \cap \{y>0\}; \\ \mathbb{C}E(u)(0,y)[\nu] = -\mathbb{C}E(u)(b,y)[\nu] & \text{for } 0<y<h(0)=h(b); \\ k + Q(E(u)) = \text{const} & \text{on } \Gamma_h \cap \{y>0\}, \end{cases} \tag{7}$$

where ν denotes the outer unit normal to Ω_h and k is the curvature of Γ_h. Due to (3), equation $(7)_1$ is a linear elliptic system satisfying the Legendre-Hadamard condition.

Definition 3. *Let $(h,u) \in X(e_0;b)$ be such that $h \in C^2([0,b])$. We say that the pair (h,u) is a critical point for F if it satisfies* (7).

Before stating the regularity result, we need to introduce the set of *cusp points* of a function $g \in AP(0,b)$

$$\Sigma_{g,c} := \{(x,g(x)) : x \in [0,b)\,, g^-(x) = g(x)\,, \text{ and } g'_+(x) = -g'_-(x) = +\infty\}\,,$$

where g^- is defined in (1), while g'_+ and g'_- denote the right and left derivatives, respectively. As before, the set $\Sigma^{\#}_{g,c}$ is obtained by replacing $[0,b)$ by $\mathbb{R}$ in the previous formula and coincides with the b-periodic extension of $\Sigma_{g,c}$.

Theorem 4 (Regularity of local minimizers). *Let $(h,u) \in X(e_0;b)$ be a b-periodic local minimizer for F. Then the following regularity results hold:*

(i) cusp points and vertical cracks are at most finite in $[0,b)$, i.e.,

$$\operatorname{card}\left(\{x \in [0,b) : (x,y) \in \Sigma_h \cup \Sigma_{h,c} \text{ for some } y \geq 0\}\right) < +\infty\,;$$

(ii) the curve $\Gamma^{\#}_h$ is of class C^1 away from $\Sigma^{\#}_h \cup \Sigma^{\#}_{h,c}$ and

$$\lim_{x \to x_0^{\pm}} h'(x) = \pm\infty \qquad \text{for every } x_0 \in \Sigma^{\#}_h \cup \Sigma^{\#}_{h,c};$$

(iii) $\Gamma^{\#}_h \cap \{(x,y) : y>0\}$ is of class $C^{1,\alpha}$ away from $\Sigma^{\#}_h \cup \Sigma^{\#}_{h,c}$ for all $\alpha \in (0,1/2)$;

(iv) let $A := \{x \in \mathbb{R} : h(x) > 0$ and h is continuous at $x\}$. Then A is an open set of full measure in $\{h>0\}$ and h is analytic in A.

Statement (ii) of Theorem 4 implies in particular that the *zero contact angle condition* between film and substrate holds. On the other hand, if $h>0$, Γ_h is of class $C^{1,\alpha}$ for all $\alpha \in (0,1/2)$, and $(h,u) \in X(e_0;b)$ satisfies the first three equations in (7), then the elliptic regularity (see [9, Proposition 8.9]) implies that $u \in C^{1,\alpha}(\overline{\Omega}_h)$ for all $\alpha \in (0,1/2)$. Moreover, if also $(7)_4$ holds in the

distributional sense, then the results contained in [14, Subsection 4.2] imply that (h,u) is analytic.

The proof of Theorem 4 is quite long. We describe here the principal steps and the main ideas. For the details we refer to [8].

The first step consists in removing the constraint $|\Omega_h| = d$ by showing that if (h,u) is a b-periodic local minimizer, then (h,u) is also a local minimizer of the penalized functional

$$(g,v) \in X(e_0;b) \mapsto F(g,v) + \Lambda\big||\Omega_g| - d\big|\,,$$

for some $\Lambda > 0$ sufficiently large. This gives a much larger choice of variations and in particular allows us to prove that $\Omega_h^\#$ satisfies an interior uniform ball condition, namely that if $\varrho > 0$ is sufficiently small (depending on u), then for all $z_0 \in \Gamma_h^\#$ there exists an open ball $B_\varrho(z) \subset \Omega_h^\#$ such that $\partial B_\varrho(z) \cap \Gamma_h^\# = \{z_0\}$.

In fact, suppose on the contrary that there exists a ball $B_\varrho(z) \subset \Omega_h^\#$ whose boundary touches $\Gamma_h^\#$ at two points (x_1,y_1), (x_2,y_2). To fix the ideas, and with no loss of generality, let us assume that $0 \le x_1 < x_2 < b$. Let us denote by $\tilde h$ the function coinciding with h in $[0,b) \setminus [x_1,x_2]$ and defined in (x_1,x_2) as the affine function whose graph is the segment connecting (x_1,y_1) and (x_2,y_2). Notice that if (h,u) satisfies the local minimality condition (6) for some $\delta > 0$ there exists $\varrho_0 > 0$ such that if $0 < \varrho < \varrho_0$, then $d_H(\Gamma_h \cup \Sigma_h, \Gamma_{\tilde h} \cup \Sigma_{\tilde h}) < \delta$. A simple calculation then shows that

$$\big[F(h,u) + \Lambda\big||\Omega_h| - d\big|\big] - \big[F(\tilde h,u) + \Lambda\big||\Omega_{\tilde h}| - d\big|\big] \ge (L-\ell) - \Lambda|D|\,, \qquad (8)$$

where ℓ is the length of the segment joining (x_1,y_1) and (x_2,y_2), L the length of the arc in Γ_h connecting the same two points, and D is the region bounded by this arc and the segment. Since by the isoperimetric inequality

$$L - \ell \ge \frac{\kappa}{\varrho}|D|\,, \qquad (9)$$

for some universal constant $\kappa > 0$, one gets that if $\varrho < \kappa/\Lambda$ the right hand side of (8) is positive, thus contradicting the local minimality of (h,u).

As a consequence of the interior ball condition one has that $\Gamma_h^\#$ has (locally) finitely many vertical cuts and cusp points. Moreover, outside these singular points $\Gamma_h^\#$ is the union of (locally) finitely many graphs of Lipschitz functions having right and left derivatives at each point, right and left continuous, respectively (see [5]). To prove that no such corner points exist and thus that $\Gamma_h^\# \setminus \big(\Sigma_h^\# \cup \Sigma_{h,c}^\#\big)$ is the union of (locally) finitely many C^1 arcs, one argues again by contradiction. In fact, if $z_0 = (x_0,y_0)$ is a corner point, then a blow-up argument, combined with the classical results due to Grisvard ([13]) on the

singularities at corner points of solutions to elasticity systems, gives that there exist $r_0, C_0 > 0$ such that for all $0 < r \le r_0$

$$\int_{B_r(z_0)\cap\Omega_h} |Du|^2\, dz \le C_0 r^{2\beta}\,,$$

for some $\beta > 1/2$. Then one can extend u to the whole ball $B_{r_0}(z_0)$ in such a way that the resulting function $\tilde{u}$ satisfies for all $r < r_0$

$$\int_{B_r(z_0)} |D\tilde{u}|^2\, dz \le C_1 r^{2\beta}\,, \tag{10}$$

for some C_1 independent of r. Then, given r sufficiently small, let $(x_1, y_1), (x_2, y_2) \in \Gamma_h^{\#} \cap \partial B_r(z_0)$ be two points such that $x_1 < x_0 < x_2$ and $\Gamma_h^{\#} \cap \big((x_1, x_2) \times \mathbb{R}\big) \subset B_r(z_0)$. Defining $\tilde{h}$ as above and comparing the energies at the two pairs (h, u), $(\tilde{h}, \tilde{u})$ one easily gets from (10) and the fact that $\beta > 1/2$

$$\big[F(h,u) + \Lambda\big|\,|\Omega_h| - d\,\big|\big] - \big[F(\tilde{h},\tilde{u}) + \Lambda\big|\,|\Omega_{\tilde{h}}| - d\,\big|\big] \ge 2r(1 - \sin(\vartheta_0/2)) + o(r)\,,$$

where ϑ_0 is the angle at the corner point z_0. Hence, the local minimality of (h, u) implies that $\vartheta_0 = \pi$ and thus that there is no corner at z_0.

The proof of statement (iii) in Theorem 4 combines in a similar way elliptic regularity and variational arguments, while (iv) follows from the regularity results proved in a more general framework in [14].

3. Second Variation and Minimality

In this section we present some qualitative and quantitative results dealing with the local and global minimality of the flat configuration. We state also a few theorems concerning the non occurrence of singularities in non flat global minimizers. The results contained in this section were all proved by the authors in [9]. Ultimately, they all rely on a new sufficient condition for local minimality, stated in Theorem 6, which provides the first extension of the classical criteria based on the positivity of second variation to the context of functionals with bulk and surface energies.

3.1. The second variation. Let $(h, u) \in X(e_0; b)$ be an admissible configuration such that $h \in C^\infty(\mathbb{R})$ is strictly positive and let $\varphi \in \widetilde{H}^1_{\#}(0, b)$, where

$$\widetilde{H}^1_{\#}(0,b) := \left\{ \varphi \in H^1(0,b) :\, \varphi(0) = \varphi(b),\ \int_0^b \varphi\, dx = 0 \right\}.$$

For $t \in \mathbb{R}$ sufficiently small, we set $h_t := h + t\varphi$ and we let u_t be the elastic equilibrium corresponding to Ω_{h_t} under the usual periodicity and boundary

conditions; i.e., $(h_t, u_t) \in X(e_0; b)$ and

$$\int_{\Omega_{h_t}} \mathbb{C}E(u_t) : E(w)\, dz = 0 \qquad \text{for all } w \in A(\Omega_{h_t}),$$

where for $g \in AP(0, b)$

$$A(\Omega_g) := \{w \in LD_{\#}(\Omega_g; \mathbb{R}^2) : \, w(\cdot, 0) \equiv 0\}\,.$$

We define the *second variation of F at (h, u) along the direction φ* as

$$\partial^2 F(h, u)[\varphi] := \frac{d^2}{dt^2} F(h_t, u_t)|_{t=0}\,.$$

The first result, also proved in [9], gives a representation formula for the second variation. To this aim, it is convenient to associate to each function $\varphi : (0, b) \to \mathbb{R}$ the corresponding *lifting* $\tilde{\varphi}$ to Γ_h by setting

$$\tilde{\varphi}(x, y) := \frac{\varphi(x)}{\sqrt{1 + h'^2(x)}} \qquad \text{for all } (x, y) \in \Gamma_h.$$

Theorem 5 (Second variation formula). *Let $h \in C^\infty(\mathbb{R})$ be a strictly positive, b-periodic function and u the corresponding elastic equilibrium. For all $\varphi \in \widetilde{H}^1_{\#}(0, b)$ we have*

$$\begin{aligned} \partial^2 F(h, u)[\varphi] = -2 \int_{\Omega_h} Q(E(v_\varphi))\, dz + \int_{\Gamma_h} (\partial_\tau \tilde{\varphi})^2 d\mathcal{H}^1 \\ + \int_{\Gamma_h} \big(\partial_\nu [Q(E(u)] - k^2\big)\tilde{\varphi}^2 d\mathcal{H}^1 - \int_{\Gamma_h} \big(Q(E(u)) + k\big)\partial_\tau (h'\tilde{\varphi}^2) d\mathcal{H}^1\,, \end{aligned} \tag{11}$$

where k is the curvature of Γ_h and v_φ is the unique solution in $A(\Omega_h)$ to the linear system

$$\int_{\Omega_h} \mathbb{C}E(v_\varphi) : E(w)\, dz = \int_{\Gamma_h} \mathrm{div}_\tau\big(\tilde{\varphi}\mathbb{C}E(u)\big) \cdot w\, d\mathcal{H}^1 \qquad \forall\, w \in A(\Omega_h)\,.$$

In (11) we have denoted by $\partial_\tau, \partial_\nu$ the tangential and normal derivative, respectively. Notice that if (h, u) is a critical point for F, i.e., a solution to the Euler–Lagrange system, the last equation in (7) implies that the last integral in (11) is zero. Even in this case, the formula representing the second variation is quite involved.

We can now state the following result, relating the positiveness of the second variation to the local minimality.

Theorem 6 (Local minimality criterion). *Let $(h, u) \in X(e_0; b)$ be a critical point for F, with $h \in C^\infty(\mathbb{R})$ and $h > 0$, and assume that the second variation of F at (h, u) is positive definite, i.e.*

$$\partial^2 F(h, u)[\varphi] > 0 \qquad \text{for all } \varphi \in \widetilde{H}^1_{\#}(0, b),\ \varphi \not\equiv 0.$$

Then (h, u) is an isolated b-periodic local minimizer in the sense of Definition 2.

Note that the regularity assumption on h is not so restrictive thanks to the remarks following the statement of Theorem 4 in the previous section.

Since this theorem is the main result of the paper and its proof is rather complicated, we outline here the overall strategy, referring to our paper [9] for all the details.

A first crucial step consists in showing that the positivity of $\partial^2 F(h,u)$ implies that (h,u) is a strict local minimizer with respect to $W^{2,\infty}$-perturbations of the profile. The proof of this minimality property follows some ideas introduced in [4] to study a similar notion of second variation for the Mumford-Shah functional. However, the presence of the vectorial elastic energy in place of the scalar Dirichlet functional requires a much more involved argument. Due to the expression of $\partial^2 F(h,u)$ given in (11), the analysis requires delicate regularity estimates in the fractional Sobolev space $H^{-\frac{1}{2}}$ of the traces of the gradient of $E(u)$ on Γ_h which were not available in the literature.

The remaining part of the proof of Theorem 6 is devoted to showing that the $W^{2,\infty}$-local minimality is in fact equivalent to the local minimality with respect to any admissible profile sufficiently close in the sense of Definition 2. Assume by contradiction that the $W^{2,\infty}$-local minimizer (h,u) is not a local minimizer. Then one can find a sequence of configurations (k_n, w_n) with $d_H\big(\Gamma_h, \Gamma_{k_n} \cup \Sigma_{k_n}\big) \le \frac{1}{n}$, $|\Omega_{k_n}| = |\Omega_h|$, and $F(k_n,w_n) \le F(h,u)$. We consider the obstacle problems

$$\min\big\{F(g,v) + \Lambda\big||\Omega_g| - |\Omega_h|\big| : (g,v) \in X,\ g \ge h - \tfrac{1}{n}\big\}\,, \tag{12}$$

with $\Lambda > 0$, and we let (g_n, v_n) be the corresponding minimizing configurations. Notice that we have replaced the volume constraint by a penalization term. Since (k_n, w_n) is an admissible competitor for (12), we have in particular

$$F(g_n,v_n) \le F(g_n,v_n) + \Lambda\big||\Omega_{g_n}| - |\Omega_h|\big| \le F(k_n,w_n) \le F(h,u)\,. \tag{13}$$

We conclude by showing that if $\Lambda > \Lambda_0 \equiv \Lambda_0(\mu,\lambda,e_0)$, then g_n is regular and $g_n \to h$ in $W^{2,\infty}$, which together with (13), gives a contradiction to the $W^{2,\infty}$-local minimality of (h,u).

The proof of the regularity and convergence of g_n is obtained by refining in a quantitative fashion the regularity estimates for minimal configurations proved in [8] and outlined at the end of the previous section. The argument goes as follows. We first show that if Λ is sufficiently large, then (h,u) is the unique minimizer to

$$\min\big\{F(g,v) + \Lambda\big||\Omega_g| - |\Omega_h|\big| : (g,v) \in X,\ g \ge h\big\}\,.$$

From this fact we deduce that (g_n, v_n) must converge (in a suitable sense) to (h,u). In particular, one can show that

$$g_n \to h \quad \text{in } L^\infty(0,b)\,. \tag{14}$$

Next, we observe that from the representation formula (4) of F the profile g_n minimizes the functional

$$g \to \mathcal{H}^1(\Gamma_g) + 2\mathcal{H}^1(\Sigma_g) + \Lambda\big||\Omega_g| - |\Omega_h|\big|$$

among all admissible g such that $h - \frac{1}{n} \le g \le g_n$. This one-sided minimality property alone suffices to provide a lower bound for the curvature (in a generalized sense) of $\Gamma_{g_n} \cup \Sigma_{g_n}$. More precisely, using the isoperimetric inequality (9) we show that for all $z_0 \in \Gamma_{g_n} \cup \Sigma_{g_n}$ there exists a ball $B_{\rho_0}(z) \subset \Omega^{\#}_{g_n}$ such that $z_0 \in \partial B_{\rho_0}(z)$, with $\rho_0 \equiv \rho_0(\Lambda)$ independent of n. As a purely geometric consequence of this uniform inner ball condition and of (14), we deduce that g_n has no cusps nor vertical cut for n large and, in fact, $g_n \to h$ in $C^1([0,b])$.

Exploiting this last convergence we obtain, as in the proof of Theorem 4, that for all $\beta \in (0,1)$

$$\int_{B_r(z_0)\cap\Omega_{g_n}} |\nabla v_n|^2 \, dz \le C_0 r^{2\beta}$$

for all $z_0 \in \Gamma_{g_n}$, $r \in (0, r_0)$, where C_0 and r_0 are independent of n. With this estimate at hand, a comparison argument similar to the one outlined at the end of Section 2 implies a uniform bound of the $C^{1,\alpha}$-norms of $\{g_n\}$ for all $\alpha \in (0, \frac{1}{2})$ and, in turn, by elliptic regularity, of $\{v_n\}$. This allows us to use the Euler-Lagrange equations to finally deduce the desired $W^{2,\infty}$-convergence of g_n to h.

Using Theorem 5 we can now calculate the second variation of the flat configuration. Given $d > 0$, the *flat configuration with volume* d is the pair $(d/b, u_{e_0}) \in X(e_0; b)$, where

$$u_{e_0}(x,y) := e_0\left(x, \frac{-\lambda}{2\mu+\lambda} y\right).$$

Note that $(d/b, u_{e_0})$ is a critical point for the functional F; i.e., it satisfies (7).

Let us fix $\varphi \in \widetilde{H}^1_{\#}(0,b)$ and set $R = (0,b) \times (0, d/b)$. From (11) we have

$$\partial^2 F\big(d/b, u_{e_0}\big)[\varphi] = -2\int_R Q(E(v_\varphi))\, dz + \int_0^b \varphi'^2 \, dx\,,$$

where v_φ is the solution in $A(R)$ of the system

$$\begin{cases} (2\mu+\lambda)\dfrac{\partial^2 v^1_\varphi}{\partial x^2} + \mu\dfrac{\partial^2 v^1_\varphi}{\partial y^2} + (\lambda+\mu)\dfrac{\partial^2 v^2_\varphi}{\partial x \partial y} = 0 \quad \text{in } R, \\ \mu\dfrac{\partial^2 v^2_\varphi}{\partial x^2} + (2\mu+\lambda)\dfrac{\partial^2 v^2_\varphi}{\partial y^2} + (\lambda+\mu)\dfrac{\partial^2 v^1_\varphi}{\partial x \partial y} = 0 \quad \text{in } R, \end{cases}$$

satisfying the usual periodicity assumptions and the boundary conditions

$$\begin{cases} \dfrac{\partial v_\varphi^1}{\partial y}+\dfrac{\partial v_\varphi^2}{\partial x} = \dfrac{4(\mu+\lambda)e_0}{2\mu+\lambda}\varphi', \ \lambda\dfrac{\partial v_\varphi^1}{\partial x}+(2\mu+\lambda)\dfrac{\partial v_\varphi^2}{\partial y} = 0 & \text{on } \{y=d/b\}, \\ v_\varphi = 0 & \text{on } \{y=0\}\,. \end{cases}$$

The above system can be explicitely solved (see [12] and [9]), thus leading to an explicit formula for $\partial^2 F\big(d/b, u_{e_0}\big)$. To this aim, we set

$$\nu_p := \frac{\lambda}{2(\lambda+\mu)}, \qquad\qquad \tau := e_0\frac{4\mu(\mu+\lambda)}{2\mu+\lambda}$$

and introduce the function J defined for $y \geq 0$ as

$$J(y) := \frac{y+(3-4\nu_p)\sinh y\cosh y}{4(1-\nu_p)^2+y^2+(3-4\nu_p)\sinh^2 y}\,.$$

The quantity ν_p is often called the *Poisson modulus* of the elastic material.

Proposition 7. *Given $d>0$, for all $\varphi \in \widetilde{H}^1_{\#}(0,b)$ one has*

$$\partial^2 F\big(d/b, u_{e_0}\big)[\varphi] = \sum_{n\in\mathbb{Z}} n^2\varphi_n\varphi_{-n}\left[1-\frac{\tau^2(1-\nu_p)bJ(2\pi nd/b^2)}{2\pi\mu n}\right],$$

where the φ_n's are the Fourier coefficients of φ in the interval $(0,b)$.

3.2. Local and global minimizers.

Combining the minimality criterion Theorem 6 with the explicit expression of the second variation, we immediately obtain the sharp necessary and suffcient conditions for the local minimality stated in Theorem 8 below. Before that we introduce the *Grinfeld function* K defined for $y \geq 0$ by

$$K(y) := \max_{n\in\mathbb{N}} \frac{1}{n} J(ny)\,.$$

It turns out (see [9, Corollary 5.3]) that,

$$K \text{ is strictly increasing and continuous, } K(y) \leq Cy, \text{ and } \lim_{y\to+\infty} K(y) = 1\,,$$

for some positive constant C.

Theorem 8 (Local minimality of the flat configuration). *Let $d_{\mathrm{loc}} : (0,+\infty) \to (0,+\infty]$ be defined as $d_{\mathrm{loc}}(b) := +\infty$, if $0 < b \leq \frac{\pi}{4}\frac{2\mu+\lambda}{e_0^2\mu(\mu+\lambda)}$, and as the solution to*

$$K\left(\frac{2\pi d_{\mathrm{loc}}(b)}{b^2}\right) = \frac{\pi}{4}\frac{2\mu+\lambda}{e_0^2\mu(\mu+\lambda)}\frac{1}{b}\,,$$

otherwise. Then the flat configuration $(d/b, u_{e_0})$ is an isolated b-periodic local minimizer for F if $0 < d < d_{\mathrm{loc}}(b)$.

The threshold $d_{\rm loc}$ is critical: indeed, for $d > d_{\rm loc}(b)$ there exists $(g, v) \in X(e_0; b)$, with $|\Omega_g| = d$, and $d_H(\Gamma_g \cup \Sigma_g, \Gamma_{d/b})$ arbitrarily small such that $F(g, v) < F(d/b, u_{e_0})$.

We come now to the issue of the global minimality of the flat configuration. Next result shows that given $b > 0$, the flat configuration is the unique b-periodic global minimizer provided the thickness d/b is small enough and that if b is sufficiently small then the flat configuration is the unique b-periodic global minimizer no matter how large the thickness of the film is.

Theorem 9 (Global minimality of the flat configuration). *The following two statements hold.*

(i) For every $b > 0$, there exists $0 < d_{\rm glob}(b) \le d_{\rm loc}(b)$ (see Theorem 8) such that the flat configuration $(d/b, u_{e_0})$ is a b-periodic global minimizer if and only $0 < d \le d_{\rm glob}(b)$. Moreover, if $0 < d < d_{\rm glob}(b)$, then $(d/b, u_{e_0})$ is the unique b-periodic global minimizer.

(ii) There exists $0 < b_{\rm crit} \le \frac{\pi}{4} \frac{2\mu+\lambda}{e_0^2 \mu(\mu+\lambda)}$ such that $d_{\rm glob}(b) = +\infty$ if and only if $0 < b \le b_{\rm crit}$, i.e., the flat configuration $(d/b, u_{e_0})$ is the unique b-periodic global minimizer for all $d > 0$ if and only if $0 < b \le b_{\rm crit}$.

The results of Theorem 9 are more qualitative than those of Theorem 8. In particular, the function $d_{\rm glob}$ and the constant $b_{\rm crit}$ are not analitycally determined and it is an open problem to establish whether or not $b_{\rm crit} < \frac{\pi}{4} \frac{2\mu+\lambda}{e_0^2 \mu(\mu+\lambda)}$ and $d_{\rm glob}(b) < d_{\rm loc}(b)$. However, next result shows that the latter inequality holds, at least for b large.

Proposition 10 (Comparison between local and global minimality thresholds). *There exists a constant $c_0 \equiv c_0(\lambda, \mu)$ such that*

$$\frac{d_{\rm loc}(b)}{b} \ge \frac{c_0}{e_0^2} \qquad \text{for all } b > 0\,.$$

Moreover,

$$\lim_{b \to \infty} \frac{d_{\rm glob}(b)}{b} = 0\,.$$

As a consequence of the previous proposition, we may prove a non-uniqueness result.

Theorem 11 (Non uniqueness). *Let $b > 0$ such that $d_{\rm glob}(b) < d_{\rm loc}(b)$. Then the minimum problem (5) with $d = d_{\rm glob}(b)$ has at least another solution besides the flat configuration $(d_{\rm glob}(b)/b, u_{e_0})$.*

We next address the occurrence of regular non flat minimal configurations. The following theorem gives an analytical confirmation of the numerical and experimental observations that singularities do not form when the sample is not too large in width and thickness.

Theorem 12 (Regular non-flat minimal configurations). *Let $b_{\rm crit}$ be the constant introduced in Theorem 9. Then the following two statements hold.*

(i) If $b_{\rm crit} < b < \frac{2\mu+\lambda}{e_0^2\mu(\mu+\lambda)}$, then for every b-periodic non-flat global minimizer (h,u) we have $h \in C^1([0,b])$.

(ii) Assume $\lambda \geq -\frac{17}{18}\mu$. There exist $b_{\rm reg} > \frac{2\mu+\lambda}{e_0^2\mu(\mu+\lambda)}$ and $d_0 > 0$ with the following property: If $\frac{2\mu+\lambda}{e_0^2\mu(\mu+\lambda)} \leq b < b_{\rm reg}$ and $d_{\rm glob}(b) \leq d < d_{\rm glob}(b)+d_0$, then for every b-periodic global minimizer (h,u) with $|\Omega_h| = d$ we have $h \in C^1([0,b])$.

In both cases (h,u) satisfies all the conclusions of Theorem 4, with $\Sigma_h^\# = \Sigma_{h,c}^\# = \emptyset$.

The last result deals with the existence of nontrivial analytic minimal configurations. It states that if b is small enough, then b-periodic non-flat global minimizers are analytic.

Theorem 13 (Analytic non-flat minimal configurations). *Let $b_{\rm crit}$ be the constant introduced in Theorem 9. There exists $\eta_0 > 0$ such that if $b = b_{\rm crit} + \eta$, with $\eta \in (0,\eta_0)$, and $(h,u) \in X(e_0;0,b)$ is any non-flat b-periodic global minimizer, then (h,u) is analytic; more precisely, h is strictly positive and analytic over $\mathbb{R}$ and, in turn, u is analytic in $\overline{\Omega}_h^\#$.*

References

[1] R.J. Asaro, W.A.Tiller, *Interface morphology development during stress corrosion cracking: Part I: Via surface diffusion.* Metall. Trans. **3** (1972), 1789–1796.

[2] E. Bonnetier, A. Chambolle, *Computing the equilibrium configuration of epitaxially strained crystalline films.* SIAM J. Appl. Math. **62** (2002), 1093–1121.

[3] E. Bonnetier E., R.S. Falk R.S., M.A. Grinfeld, *Analysis of a one-dimensional variational model of the equilibrium shape of a deformable crystal.* M2AN Math. Model. Numer. Anal. **33** (1999), 573–591.

[4] F. Cagnetti, M.G. Mora, M. Morini, *A second order minimality condition for the Mumford-Shah functional.* Calc. Var. Partial Differential Equations **33** (2008), 37–74.

[5] A. Chambolle, C.J. Larsen, *C^∞ regularity of the free boundary for a two-dimensional optimal compliance problem.* Calc. Var. Partial Differential Equations **18** (2003), 77–94.

[6] A. Chambolle, M. Solci, *Interaction of a bulk and a surface energy with a geometrical constraint.* SIAM J. Appl. Math. **39** (2007), 77–102.

[7] I.Fonseca, N. Fusco, G. Leoni, V. Millot, *Material voids for anisotropic surface energies.* Work in progress.

[8] I.Fonseca, N. Fusco, G. Leoni, M. Morini, *Equilibrium configurations of epitaxially strained crystalline films: existence and regularity results.* Arch. Rational Mech. Anal. **186** (2007), 477–537.

[9] N. Fusco, M. Morini, *Equilibrium configurations of epitaxially strained elastic films: second order minimality conditions and qualitative properties of solutions.* Preprint 2009. Downloadable from the preprint server *http://cvgmt.sns.it*

[10] H. Gao, W.D. Nix, *Surface roughening of heteroepitaxial thin films.* Annu. Rev. Mater. Sci. **29** (1999), 173–209.

[11] M.A. Grinfeld, *Instability of the separation boundary between a non-hydrostatically stressed elastic body and a melt.* Soviet Physics Doklady **31** (1986), 831–834.

[12] M.A. Grinfeld, *Stress driven instabilities in crystals: mathematical models and physical manifestation.* J. Nonlinear Sci. **3** (1993), 35–83.

[13] P. Grisvard, *Singularités en elasticité.* Arch. Rational Mech. Anal. **107** (1989), 157–180.

[14] H. Koch, G. Leoni, M. Morini, *On Optimal regularity of Free Boundary Problems and a Conjecture of De Giorgi.* Comm. Pure Applied Math. **58** (2005), 1051–1076.

[15] B.J. Spencer, *Asymptotic derivation of the glued-wetting-layer model and contact-angle condition for Stranski-Krastanow islands.* Physical Review B **59** (1999), no. 3, 2011–2017.

[16] B.J. Spencer, D.I. Meiron, *Nonlinear evolution of stress-driven morphological instability in a two-dimensional semi-infinite solid.* Acta Metall. Mater. **42** (1994), 3629–3641.

[17] B.J. Spencer, J. Tersoff, *Equilibrium shapes and properties of epitaxially strained islands.* Physical Review Letters **79** (1997), 4858–4861.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Weak Solutions of Nonvariational Elliptic Equations

Nikolai Nadirashvili* and Serge Vlăduţ†

Abstract

We discuss basic properties (uniqueness and regularity) of viscosity solutions to fully nonlinear elliptic equations of the form $F(x, D^2u) = 0$, which includes also linear elliptic equations of nondivergent form. In the linear case we consider equations with discontinuous coefficients.

Mathematics Subject Classification (2010). Primary 35J15, 35D30, 35D40, 35J60 Secondary 17A35, 20G41, 53C38, 60G46 .

Keywords. Fully nonlinear elliptic equations, viscosity solutions, stochastic processes, triality, division algebras, Hessian equations, Isaacs equation, special Lagrangian equation

1. Introduction

We consider elliptic equations written in terms of the Hessian D^2u of the unknown function u, i.e. we consider equations of the form

$$F(D^2u(x), x) = 0, \tag{1.1}$$

where u is a C^2-function defined on a domain of $\mathbf{R}^n$.

The equation (1.1) include as principal cases the fully nonlinear equation depending only on the Hessian,

$$F(D^2u(x)) = 0, \tag{1.2}$$

*Laboratoire d'Analyse, Topologie, Probabilité, Centre de Mathématiques et Informatique, 39, rue F. Joliot Curie, 13453 Marseille Cedex 13, France. E-mail: nicolas@cmi.univ-mrs.fr.

†Institut de Mathématiques de Luminy, Campus de Luminy, Case 907, 13288 Marseille Cedex 9, France. E-mail: vladut@iml.univ-mrs.fr.

and the linear equation,

$$\sum a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} = 0. \tag{1.3}$$

The ellipticity conditions conditions for the equations (1.2) and (1.3) can be written in the following form. We assume that F is a Lipschitz function defined on an open set $D \subset S^2(\mathbf{R}^n)$ of the space of $n \times n$ symmetric matrices. The equation (1.2) is called uniformly elliptic if there exists a constant $C = C(F) \geq 1$ (called an *ellipticity constant*) such that

$$C^{-1}||N|| \leq F(M+N) - F(M) \leq C||N|| \tag{1.4}$$

for any non-negative definite symmetric matrix N; if $F \in C^1(D)$ then this condition is equivalent to

$$\frac{1}{C}|\xi|^2 \leq F_{u_{ij}}\xi_i\xi_j \leq C|\xi|^2 \ , \forall \xi \in \mathbf{R}^n. \tag{1.5}$$

Here, u_{ij} denotes the partial derivative $\partial^2 u/\partial x_i \partial x_j$. The equation (1.2) is called stritly elliptic if simply $0 < F(M+N) - F(M)$ for any positive definite symmetric matrix N; if $F \in C^1(D)$ then this is equivalent to $0 < F_{u_{ij}}$. A function u is called a *classical* solution of (1.1) if $u \in C^2(\Omega)$ and u satisfies (1.1). Actually, any classical solution of (1.1) is a smooth ($C^{\alpha+3}$) solution, provided that F is a smooth (C^α) function of its arguments.

For the linear equation (1.3) we assume that the coefficients $a_{ij}(x)$ are measurable functions which satisfy the inequalities

$$C^{-1}|\xi|^2 \leq \sum a_{ij}\xi_i\xi_j \leq C|\xi|^2,$$

where $C > 0$ is an ellipticity constant.

Solutions of equations (1.2) and (1.3) have an important connection. Let u_1, u_2 be two classical solutions of the equation (1.2). Then the difference $u = u_1 - u_2$ is a solution of a linear uniformly elliptic equation (1.3):

$$\sum a_{ij}(x)\frac{\partial^2 (u_1 - u_2)}{\partial x_i \partial x_j} = 0. \tag{1.6}$$

The most of concrete examples of fully nonlinear equation (1.2) are invariant under the rotations of the Euclidean space $\mathbf{R}^n$, i.e., function F is invariant under the action of the group $O(n)$ on $S^2(\mathbf{R}^n)$:

$$\forall O \in O(n),\ F({}^tO \cdot S \cdot O) = F(S)\ . \tag{1.7}$$

Such equations are called *Hessian equations*. In other words, denote by

$$\lambda(S) = \{\lambda_i : \lambda_1 \leq ... \leq \lambda_n\} \in \mathbf{R}^n$$

the (ordered) set of eigenvalues of the matrix $S \in S^2(\mathbf{R}^n)$. Equation (1.2) is called a Hessian equation if the function $F(S)$ depends only on the eigenvalues $\lambda(S)$ of the matrix S, i.e., if

$$F(S) = f(\lambda(S)),$$

for some function f on $\mathbf{R}^n$ invariant under permutations of the coordinates.

If we assume that the function $F(S)$ is defined for any symmetric matrix S, i.e., $D = S^2(\mathbf{R}^n)$ the Hessian invariance relation (1.7) implies the following:

(a) F is a smooth (real-analytic) function of its arguments if and only if f is a smooth (real-analytic) function.

(b) Inequalities (1.4) are equivalent to the inequalities

$$\frac{\mu}{C_0} \leq f(\lambda_i + \mu) - f(\lambda_i) \leq C_0\mu, \ \forall \mu \geq 0,$$

$\forall i = 1, ..., n$, for some positive constant C_0.

(c) F is a concave function if and only if f is concave.

Well known examples of the fully nonlinear equations are Laplace, Monge-Ampère and Special Lagrangian equation (which are Hessian equations), Bellman and Isaacs equations.

Bellman and Isaacs equations appear in the theory of controlled diffusion processes. The both are fully nonlinear uniformly elliptic equations of the form (1.1). The Bellman equation is concave in $D^2u \in S^2(\mathbf{R}^n)$ variables. However, Isaacs operators are, in general, neither concave nor convex. In a simple homogeneous form the Isaacs equation can be written as follows:

$$F(D^2u) = \sup_b \inf_a L_{ab}u = 0, \tag{1.8}$$

where L_{ab} is a family of linear uniformly elliptic operators of type (1.3) with an ellipticity constant $C > 0$ which depends on two parameters a, b.

The important difference of the equation (1.1) from the equations with the origin in the calculus of variations is that the last ones have a divergent structure which allows to use integral identities to define weak solutions of the equations. For the equations (1.2) and (1.3) weak extension of the solutions known as the *viscosity* solutions can be done in different ways.

For the fully nonlinear equation (1.2) the set of the viscosity solution can be defined as the intersection of C-closures of the sets of classical super and subsolutions. For the Bellman and Isaacs equation the probabilistic solutions

can be defined as well, [19, 12, 24]. For the linear operator (1.3) one can define a continuous strong Markov process $x(t)$ such that $x(t+h)-x(t)$ for $h \to 0$ $x(t)$ behaves as a Gaussian process with mean zero and covariance $a(x(t))$.

One can define the viscosity solutions for the equation (1.1) with just measurable dependence of the function F on x, see [7], the case which uniforms the equations (1.2) and (1.3). However, we will consider equations (1.2) and (1.3) separately. Remarkably, the equality (1.6) holds for the viscosity extension of the classical solutions of (1.2) and (1.3). Our main goal is to discuss the basic properties of the viscosity solutions of the equations (1.2) and (1.3).

2. Stochastic Processes and Viscosity Solutions of Linear Elliptic Equations

First we recall first the formal definitions of viscosity solutions to the equation (1.3). Let $\Omega \subset \mathbf{R}^n$ be a smooth bounded domain. Let L be a linear uniformly elliptic operator (1.3) defined in Ω with the ellipticity constant C. We consider a Dirichlet problem in Ω:

$$\begin{cases} Lv = 0 & \text{in } \Omega \\ v = \varphi & \text{on } \partial\Omega \end{cases} \tag{2.1}$$

Definition 2.1. The function v is a viscosity solution of (1) if

$$v = \lim v_k,$$

where

$$Lv_k = \sum a_{ij}^k(x) \frac{\partial^2 v_k}{\partial x_i \partial x_j} = 0,$$

a_{ij}^k are continuous and $a_{ij}^k \to a_{ij}$ in $L_1(\Omega)$.

Extending the notion of sub and super solutions Jensen [17] suggested an equivalent definition of a viscosity solution of (1.3) which is close to the definition of a viscosity solution to fully nonlinear elliptic equations.

The existence of a viscosity solution to the Dirichlet problem (2.1) immediately follows from Definition 2.1. The important problem is the uniqueness of the viscosity solution of (2.1). The uniqueness of the viscosity solution is related to the uniqueness of the diffusion generated by the operator L. The diffusion (ξ_t, P_x) related to the operator L can be defined for example as a solution of the martingale problem, [36]. The diffusion (ξ_t, P_x) defines a solution of Dirichlet problem (2.1):

$$u(x) = E_x\{\varphi(\xi_\tau)\},$$

where τ is the first time when the path leaves the domain Ω.

By Krylov, [21], and Stroock, Varadhan, [36], it is known that the diffusion (ξ_t, P_x) is unique if $n = 2$ or if coefficients of L are continuous functions. Hence in these cases we have the uniqueness of the viscosity solution of the Dirichlet problem (2.1). In the general case we prove the following result [25],

Theorem 2.1. *There exists a uniformly elliptic operator L of the form (1.3) defined in the unit ball $B \subset \mathbf{R}^3$ and there is a function $\varphi \in C(\partial B)$, such that the Dirichlet problem (2.1) has at least two viscosity solutions.*

The coefficients of the operator L given by Theorem 2.1 are severely discontinuous, the set of discontinuity of a_{ij} has a complete Hausdorff dimension. If on the contrary, the set of discontinuity of a_{ij} is small, a singular point or has a small Hausdorff dimension, then the viscosity solution is unique, [8, 35]. We conjecture: if the set of discontinuity of coefficients a_{ij} has Hausdorff dimension less than 1, then the viscosity solution of the Dirichlet problem (2.1) is unique.

Other uniqueness results one can find in [5, 19].

One can rise a similar question on the uniqueness of the diffusion with reflection from the boundary. Assume that $\omega \subset \partial\Omega$ be a subdomain of the boundary. Let l be a vector field defined on ω and transversal to the boundary, $(l, n) > \delta > 0,\ |l| = 1$. Consider a problem,

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ \partial u/\partial l = 0 & \text{on } \omega \\ u = \varphi & \text{on } \partial\Omega \setminus \omega \end{cases} \tag{2.2}$$

Compactness results for the solutions of (2.2), see [22], allow to define viscosity solutions of (2.2) for any measurable field l. In dimensions $n \geq 3$ the question of the uniqueness of the viscosity solutions of (2.2) is open. For a discussion of linear and fully nonlinear oblique derivative problem see [4].

3. Nonclassical Solutions to Fully Nonlinear Elliptic Equations

Consider the following Dirichlet problem

$$\begin{cases} F(D^2u(x)) = 0 & \text{in } \Omega \\ u = \varphi & \text{on } \partial\Omega \end{cases} \tag{3.1}$$

where $\Omega \subset \mathbf{R}^n$ is a bounded domain with smooth boundary $\partial\Omega$, F is a uniformly elliptic operator (1.2) and φ is a continuous function on $\partial\Omega$.

It is not difficult to prove that the problem (3.1) has no more than one classical solution (see e.g. [13]). The basic problem is the existence of such classical solutions. Although the first systematic study of the Dirichlet problem for fully nonlinear equations was done by Bernstein at the beginning of the 20-th century (see [13]), the first complete result did not appear until 1953,

when Nirenberg proved the existence of a classical solution to problem (3.1) in dimension $n = 2$ ([33]). For $n \geq 3$, the problem of the existence of classical solutions to Dirichlet problem (3.1) remained open.

In order to get a solution to the problem (3.1) one can try to extend the notion of the classical solution of the equation (1.2). That was done recently: Crandall-Lions and Evans developed the concept of viscosity (weak) solutions of the fully nonlinear elliptic equations. As a characteristic property for such extension can be taken the maximum principle in the following form:

Let u_1, u_2 be two solutions of the following equations, $F(D^2u_1) = f_1$ in Ω and $F(D^2u_2) = f_2$ in Ω. Then for any subdomain $G \subset \Omega$ the inequalities $f_1 \leq f_2$ ($f_1 \geq f_2$) in G and $u_1 \geq u_2$ ($u_1 \leq u_2$) on ∂G imply the inequality $u_1 \geq u_2$ ($u_1 \leq u_2$) in G.

Such maximum principle holds for C^2 functions u_1, u_2. We call a continuous function u_1 a viscosity solution of $F(D^2u_1) = f_1$ if the above maximum principle holds for u_1 and all C^2-functions u_2.

It is possible to prove the existence of a viscosity solution to the Dirichlet problem (3.1) and Jensen's theorem says that the viscosity solution of the problem (3.1) is unique. For more details see [6, 9].

There are important classes of the fully nonlinear Dirichlet problems for which the viscosity solution is in fact a classical one, e.g., due to Evans-Krylov regularity theory, in the case when the function F is convex, (see [11, 6, 20]). Recently we have shown that in dimension 3 axial-symmetric viscosity solutions of uniformly elliptic Hessian equations are in fact the classical ones [31]. However, for the general F the problem of the coincidence of viscosity solutions with the classical remained open.

Our central result is the existence of nonclassical viscosity solution of (1.2) in the dimension 12. More precisely we prove

Theorem 3.1. *Let $\Omega \subset \mathbf{R}^{12}$ be the unit ball. Then there exist a smooth uniformly elliptic F and $\varphi \in C^\infty(\Omega)$ such that the Dirichlet problem* (3.1) *has no classical solution.*

We discuss the ideas underlying the result. We can try to find a singular viscosity solution of (1.2) in a form of a homogeneous order α function $w(x)$,

$$w(kx) = k^\alpha w(x),$$

defined in the unit ball in $\mathbf{R}^n$. On this way we immediately meet the following restrictions:

(1) From the regularity results [6] for the solutions of (1.2) it follows that α could be only in the range $1 + \epsilon \leq \alpha \leq 2,\ \epsilon > 0$.

(2) From the old result of A. Alexandrov, [2], it follows that $n \geq 4$. (The theorem of Alexandrov is valid for real analytic F. The corresponding result for smooth F was proved in [14].)

(3) If $w(x)$ is a viscosity solution of (1.2) in the whole space $\mathbf{R}^n$ then $\alpha = 2$ [32].

If after these discouraging remarks we still will to find a singular viscosity solution in the form of a homogeneous function we need first to transform problem into a question on implicit properties of the function w.

Let $A \in S^2(\mathbf{R}^n)$. We say that the symmetric matrix A is hyperbolic with the constant M, $A \in H_M$ if

$$\frac{1}{M} < -\frac{\lambda_1(A)}{\lambda_n(A)} < M.$$

Lemma 3.1. *Let*

$$w(tx) = t^2 w(x), x \in \mathbf{R}^n, t > 0$$

and

$$Hess\ w : S^{n-1} \to S^2(\mathbf{R}^n)$$

be a smooth embedding.

Assume (a) for all $x, y \in S^n$, $x \neq y$,

$$(D^2 w(x) - D^2 w(y)) \in H_M.$$

Then w *is a (viscosity) solution of a uniformly elliptic equation*

$$F(D^2 w) = 0$$

in $\mathbf{R}^n$.

We can get some simplification of the test of Lemma 3.1 for the function w if we reduce the class of admissible functions w. We will consider homogeneous order 2 functions w given in the following form,

$$w(x) = \frac{P(x)}{|x|},$$

P being a cubic form in $\mathbf{R}^n$. Let $a, b \in \mathbf{R}^n$, $|a|, |b| = 1$, $a \neq b$, $d = a - b$, $P_d = \partial P / \partial d$. Denote by $\lambda_1 \leq ... \leq \lambda_n$ the eigenvalues of the quadratic form P_d.

Lemma 3.2. *Assume that* $\lambda_3 < 0 < \lambda_{n-2}$ *and*

$$\frac{\lambda_1}{\lambda_3},\ \frac{\lambda_n}{\lambda_{n-2}} < 2 - \delta, \tag{3.2}$$

where $\delta > 0$. *Then*

$$(D^2 w(a) - D^2 w(b)) \in H_M,$$

where M *depends only on* δ.

Thus Lemma 3.2 links our initial problem of the existence of non-classical viscosity solutions with the study of cubic forms and their spectral properties (3.2) for various values of the parameter d.

As the first step towards the proper cubic form P we consider in $\mathbf{C}^3 = \mathbf{R}^6$ the form

$$P_6 = Re(z_1 z_2 z_3) = x_1 x_2 x_3 - x_1 y_2 y_3 - y_1 y_2 x_3 - y_1 x_2 y_3,$$

where $z_j = x_j + iy_j \in \mathbf{C}$. Then for $d \in \mathbf{R}^6$ it is not hard to prove the following inequalities:

$$\frac{\lambda_1}{\lambda_2},\ \frac{\lambda_6}{\lambda_5} \leq 2. \tag{3.3}$$

The inequalities (3.3) are obviously weaker than (3.2). First in (3.3) we have the second eigenvalue instead of the third, secondly the inequalities (3.3) are not strict. The second obstacle appears to be unexpectedly serious. One can prove the following general algebraic statement: For any cubic form P in $\mathbf{R}^n$ there exists a vector d such that if $\lambda_1 \leq ... \leq \lambda_n$ are the eigenvalues of the quadratic form P_d then $\lambda_1/\lambda_2 \geq 2$. Thus it is impossible to find a desirable cubic form just by means of Lemma 3.2. Fortunately we have the following upgrading of Lemma 3.2. Denote by Q the restriction of the quadratic form P_d on the hyperplane orthogonal to d and let $\lambda'_1 \leq ... \leq \lambda'_{n-1}$ be the eigenvalues of the quadratic form Q.

Lemma 3.3. *Assume that for some $\delta > 0$ and any $d \in S_1^{n-1}$ we have $\lambda_3 < 0 < \lambda_{n-2}$ and*

$$\frac{\lambda'_1}{\lambda_3},\ \frac{\lambda'_{n-1}}{\lambda_{n-3}} < 2 - \delta. \tag{3.4}$$

Then

$$(D^2 w(a) - D^2 w(b)) \in H_M,$$

where M depends only on δ.

The lemma is based on the fact that for any $e \in S_1^{n-1}, e \perp a$ one has $w_{ee}(a) = P_{ee}(a) - P(a)$, thus $w_{ee}(a) - w_{ee}(b) = P_{ee}(a) - P_{ee}(b) - (P(a) - P(b))$ for $e \perp a, b$ and taking the Taylor series of the difference $P(a) - P(b)$ at the point $c = \frac{a+b}{2} \perp d = a - b$ one gets M proportional to δ.

Now in order to win the third eigenvalue in inequalities (3.3) it seems natural to consider the Hamiltonian quaternions instead of the complex numbers. Thus we consider the cubic form

$$P = P_{12} = Re(q_1 q_2 q_3) \tag{3.5}$$

in $\mathbf{H}^3 = \mathbf{R}^{12}$ where $q_i \in \mathbf{H}$ are quaternions.

It is possible to prove that the cubic form (3.5) satisfies the inequalities (3.4). The ideas behind this fact will be discussed in the next section. Hence the function $P_{12}(x)/|x|$ is a non-classical viscosity solution in $\mathbf{R}^{12}$.

Since inequalities (3.4) are just a sufficient condition for the function w to be a viscosity solution of a fully nonlinear elliptic equation, it is interesting to understand the situation over the field $\mathbf{C}$, i.e., whether function $w_6 = P_6(x)/|x|$ satisfies the condition of Lemma 3.1. The careful analysis shows that the function w_6 is a solution of a fully nonlinear elliptic equation, but unfortunately a degenerate elliptic equation. More precisely, the equation rests strictly elliptic but loses the uniform ellipticity in a neighborhood of the subset of S_1^5 formed by the points with $|z_1| = |z_2| = |z_3|,\ Re(z_1 z_2 z_3) = 0$.

To explain why P_6 does not work and P_{12} does work we give in the next section a short excursion in the area of division algebras and exceptional Lie groups. That will lead us also to various extensions of Theorem 3.1.

4. Trialities, Quaternions, Octonions and Hessian Equations

As we have seen in the previous section, cubic forms for which the quadratic form P_d verifies the inequalities (3.2) or (3.3) should be rather exceptional. In fact all examples of such forms known to us come from trialities, which in turn are intimately related to division algebras and exceptional Lie groups. Let us recall some of their elementary properties [1, 3].

Duality is ubiquous in algebra; triality is similar, but subtler. For two real vector spaces V_1 and V_2, a duality is simply a nondegenerate bilinear map

$$f : V_1 \times V_2 \longrightarrow \mathbf{R}.$$

Similarly, for three real vector spaces V_1, V_2, and V_3, a triality is a trilinear map

$$t : V_1 \times V_2 \times V_3 \longrightarrow \mathbf{R}$$

that is nondegenerate in the sense that if we fix any two arguments to any nonzero values, the linear functional induced on the third vector space is nonzero. Each vector spaces V_1 has the dual vector space $V_2 = V_1^*$. Trialities are much rarer and in fact come from division algebras. Indeed, let

$$t : V_1 \times V_2 \times V_3 \longrightarrow \mathbf{R}$$

be a triality. By dualizing one gets a bilinear map

$$m : V_1 \times V_2 \longrightarrow V_3^*.$$

By the nondegeneracy of t, the three spaces V_1, V_2 and V_3^* can be identified with a single vector space, say V, which gives a product

$$m : V \times V \longrightarrow V.$$

Applying the nondegeneracy once more one sees that V is actually a division algebra. It follows from the well-known theorem by Bott-Kervaire-Milnor on non-parallelizability of spheres $S^n, n > 7$, that trialities only occur in dimensions 1, 2, 4, or 8. The one dimensional case is trivial and uninteresting. Examples of trialities in dimensions 2, 4 and 8 are given by

$$t_2 : \mathbf{C} \times \mathbf{C} \times \mathbf{C} \longrightarrow \mathbf{R},\ t_2(z_1, z_2, z_3) = Re(z_1 z_2 z_3),$$

$$t_4 : \mathbf{H} \times \mathbf{H} \times \mathbf{H} \longrightarrow \mathbf{R},\ t_4(q_1, q_2, q_3) = Re(q_1 q_2 q_3),$$

$$t_8 : \mathbf{O} \times \mathbf{O} \times \mathbf{O} \longrightarrow \mathbf{R},\ t_8(o_1, o_2, o_3) = Re((o_1 o_2)o_3) = Re(o_1(o_2 o_3)).$$

A choice of $\mathbf{R}$-bases in $\mathbf{C}, \mathbf{H}$ and $\mathbf{O}$ transforms t_2, t_4 and t_8 into the following cubic harmonic forms in 6,12 and 24 variables respectively:

$$P_6 = X_0Y_0Z_0 - X_0Y_1Z_1 - X_1Y_1Z_0 - X_1Y_0Z_1,$$

$$P_{12} = (Y_0Z_0 - Y_1Z_1 - Y_2Z_2 - Y_3Z_3)X_0 + (Y_3Z_2 - Y_0Z_1 - Y_1Z_0 - Y_2Z_3)X_1 +$$
$$(Y_1Z_3 - Y_0Z_2 - Y_2Z_0 - Y_3Z_1)X_2 + (Y_2Z_1 - Y_0Z_3 - Y_1Z_2 - Y_3Z_0)X_3,$$

$$P_{24} = (Z_0Y_0 - Z_1Y_1 - Z_2Y_2 - Z_3Y_3 - Z_4Y_4 - Z_5Y_5 - Z_6Y_6 - Z_7Y_7)X_0 +$$
$$(-Z_1Y_0 - Z_0Y_1 - Z_4Y_2 - Z_7Y_3 + Z_2Y_4 - Z_6Y_5 + Z_5Y_6 + Z_3Y_7)X_1 +$$
$$(-Z_2Y_0 + Z_4Y_1 - Z_0Y_2 - Z_5Y_3 - Z_1Y_4 + Z_3Y_5 - Z_7Y_6 + Z_6Y_7)X_2 +$$
$$(-Z_3Y_0 + Z_7Y_1 + Z_5Y_2 - Z_0Y_3 - Z_6Y_4 - Z_2Y_5 + Z_4Y_6 - Z_1Y_7)X_3 +$$
$$(-Z_4Y_0 - Z_2Y_1 + Z_1Y_2 + Z_6Y_3 - Z_0Y_4 - Z_7Y_5 - Z_3Y_6 + Z_5Y_7)X_4 +$$
$$(-Z_5Y_0 + Z_6Y_1 - Z_3Y_2 + Z_2Y_3 + Z_7Y_4 - Z_0Y_5 - Z_1Y_6 - Z_4Y_7)X_5 +$$
$$(-Z_6Y_0 - Z_5Y_1 + Z_7Y_2 - Z_4Y_3 + Z_3Y_4 + Z_1Y_5 - Z_0Y_6 - Z_2Y_7)X_6 +$$
$$(-Z_7Y_0 - Z_3Y_1 - Z_6Y_2 + Z_1Y_3 - Z_5Y_4 + Z_4Y_5 + Z_2Y_6 - Z_0Y_7)X_7.$$

The main property of those forms is as follows:

Lemma 4.1. *Let $d = (a, b, c) \in S_1^k \subset V^3, k = 5, 11$ or 23, for $V = \mathbf{C}, \mathbf{H}$ or $\mathbf{O}$. Let $m = m(d) = |a| \cdot |b| \cdot |c| \in [0, \frac{1}{3\sqrt{3}}]$, $n = n(d) = P(a, b, c)$, $\mid n \mid \leq m$ for $P = P_6, P_{12}$ or P_{24} respectively. The characteristic polynomial $CH_d(x)$ of the quadratic form $Q_d = 2P_d = D^2P(d)$ equals*

$$CH_d(x) = (x^3 - x + 2m)(x^3 - x - 2m) \text{ for } P = P_6$$

$$CH_d(x) = (x^3 - x + 2m)(x^3 - x - 2m)(x^3 - x + 2n)^2 \text{ for } P = P_{12}$$

$$CH_d(x) = (x^3 - x + 2m)(x^3 - x - 2m)(x^3 - x + 2n)^6 \text{ for } P = P_{24}.$$

To calculate the characteristic polynomial one uses the automorphism groups of the trialities which are Lie groups closely connected to exceptional ones. Note that the trialities t_2, t_4 and t_8 are *normed*, i.e.

$$\forall (v_1, v_2, v_3) \in V^3, \forall j \in \{2, 4, 8\},\ |t_j(v_1, v_2, v_3)| \leq |v_1||v_2||v_3|.$$

An automorphism of the normed triality

$$t : V_1 \times V_2 \times V_3 \longrightarrow \mathbf{R}$$

is a triple of norm preserving maps $f_i : V_i \longrightarrow V_i\ i = 1, 2, 3$ such that

$$\forall (v_1, v_2, v_3) \in V_1 \times V_2 \times V_3,\ t(f_1(v_1), f_2(v_2), f_3(v_3)) = t(v_1, v_2, v_3).$$

The automorphism groups of our normed trialities are

$$Aut(t_2) = \{(g_1, g_2, g_3) \in U(1)^3 : g_1 g_2 g_3 = 1\} \times \mathbf{Z}_2,$$

$$Aut(t_4) = Sp(1)^3/\{\pm(1, 1, 1)\},$$

$$Aut(t_8) = Spin(8),$$

where

$$U(1) = S^1 = \{u \in \mathbf{C} : |u| = 1\},\ Sp(1) = SU(2) = \{v \in \mathbf{H} : |v| = 1\},$$

$Spin(8)$ being the spinor group which sits in an exact sequence

$$1 \longrightarrow \mathbf{Z}_2 \longrightarrow Spin(8) \longrightarrow SO(8) \longrightarrow 1.$$

The polynomial $CH_d(x)$ is invariant under the action of the group $Aut(t_j)$ which permits to reduce the vector d to a very special form which, in turn, reduces the matrix of the form Q_d to a very special form permitting to calculate the characteristic polynomial.

More precisely, the action of $Aut(t_2)$ permits to assume that a, b are real for $d = (a, b, c) \in \mathbf{C}^3$, which gives the formula for $CH_d(x)$ after a short calculation (however, in this case the calculation is easy for a general d as well). In the quaternionic case we can assume applying the action of $Aut(t_4)$ that

$$d = (a, b, c) \in \mathbf{C}^3 \subset \mathbf{H}^3$$

and thus the matrix $M_{12,d}$ of the form Q_d takes the block form:

$$M_{12,d} = \begin{pmatrix} M_{6,d} & 0 \\ 0 & A_6 \end{pmatrix}$$

$M_{6,d}$ being the matrix for the case of $\mathbf{C}$, and A_6 being a simple structure matrix with the characteristic polynomial $(x^3 - x + 2n)^2$. Finally, in the octonionic case the action of $Aut(t_8)$ reduces d to

$$d = (a, b, c) \in \mathbf{H}^3 \subset \mathbf{O}^3$$

(in fact, much more is true, one can assume that $d = (a, b, c) \in \mathbf{R}\times\mathbf{C}\times\mathbf{H} \subset \mathbf{O}^3$) and the matrix $M_{24,d}$ takes the form

$$M_{24,d} = \begin{pmatrix} M_{12,d} & 0 \\ 0 & A_{12} \end{pmatrix}$$

for a simple structure matrix A_{12} with the characteristic polynomial $(x^3 - x + 2n)^4$.

We get for the spectrums $Sp(Q_{6,d}), Sp(Q_{12,d})$ and $Sp(Q_{24,d})$:

Corollary. *Define the angles $\alpha, \beta \in [0, \pi]$ by $m = \cos\alpha, n = \cos\beta$. Then*

$$Sp(Q_{6,d}) = \left\{ \frac{2}{\sqrt{3}} \cos\left(\frac{\alpha + \pi k}{3}\right) \right\}, k \in [0, 5].$$

$$Sp(Q_{12,d}) = \left\{ \frac{2}{\sqrt{3}} \cos\left(\frac{\alpha + \pi k}{3}\right), 2 \times \frac{2}{\sqrt{3}} \cos(\frac{\beta + \pi(2l+1)}{3}) \right\}, k \in [0, 5], l \in [0, 2].$$

$$Sp(Q_{24,d}) = \left\{ \frac{2}{\sqrt{3}} \cos\left(\frac{\alpha + \pi k}{3}\right), 6 \times \frac{2}{\sqrt{3}} \cos\left(\frac{\beta + \pi(2l+1)}{3}\right) \right\}, k \in [0, 5], l \in [0, 2].$$

This immediately implies for $Q_{12,d}$

Corollary. *If $\lambda_1/\lambda_3 = 2$ then*

$$CH_d(x) = \left(x - \frac{2}{\sqrt{3}}\right)^3 \left(x + \frac{2}{\sqrt{3}}\right) \left(x - \frac{1}{\sqrt{3}}\right)^2 \left(x + \frac{1}{\sqrt{3}}\right)^6,$$

and $d = v_1$; if $\lambda_{12}/\lambda_{10} = 2$ then

$$CH_d(x) = \left(x + \frac{2}{\sqrt{3}}\right)^3 \left(x - \frac{2}{\sqrt{3}}\right) \left(x + \frac{1}{\sqrt{3}}\right)^2 \left(x - \frac{1}{\sqrt{3}}\right)^6$$

and $d = v_{12}$ where v_i is the normalized eigenvector corresponding to λ_i.

Since in this extremal cases the eigenvalue λ_1 (resp. λ_{12}) is simple and since the set of d verifying those conditions is finite we get Lemma 3.3 for some $\delta > 0$; a more thorough analysis [26] permits to choose $\delta = 1/2$ which is still non-optimal.

Exceptional properties of $Q_{24,d}$ permit to ameliorate considerably Theorem 3.1, namely to consruct Hessian equations with *singular*, i.e. with unbouded second derivatives, solutions.

One begins with an appropiate version of Lemma 3.1.

Lemma 4.2. 1). *Let*

$$w(tx) = t^2 w(x), x \in \mathbf{R}^n, t > 0.$$

Assume
for all $x, y \in S^n$, *for all* $O \in O(n)$

$$(D^2w(x) - {}^tO \cdot D^2w(y) \cdot O) \in H_M \cup \{0\}.$$

Then w *is a (viscosity) solution of a Hessian uniformly elliptic equation*

$$F(D^2w) = 0$$

in $\mathbf{R}^n$.

2). *Let for some* $\delta \in]0, 1]$

$$w_\delta(tx) = t^{1+\delta} w_\delta(x), x \in \mathbf{R}^n, t > 0.$$

Assume
for all $x, y \in S^n$, *for all* $O \in O(n)$ *and for all* $K > 0$

$$(D^2w_\delta(x) - K \cdot {}^tO \cdot D^2w_\delta(y) \cdot O) \in H_M \cup \{0\}.$$

Then w_δ *is a (viscosity) solution of a Hessian uniformly elliptic equation*

$$F(D^2w) = 0$$

in $\mathbf{R}^n$.

Our analysis of the tests given by Lemma 4.2 is based on the following classical result by H.Weyl [40]:

Lemma 4.3. *Let* A, B *be two real symmetric matrices with the eigenvalues* $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ *and* $\lambda'_1 \leq \lambda'_2 \leq \ldots \leq \lambda'_n$ *respectively. Then for the eigenvalues* $\Lambda_1 \leq \Lambda_2 \leq \ldots \leq \Lambda_n$ *of the matrix* $A - B$ *we have*

$$\Lambda_n \geq \max_{i=1,\cdots,n} (\lambda_i - \lambda'_i), \;\; \Lambda_1 \leq \min_{i=1,\cdots,n} (\lambda_i - \lambda'_i).$$

Since for $e \in S^{23}_1, e \perp a$ one has $w_{ee}(a) = P_{ee}(a) - P(a)$, the restrition of the form $D^2w(a) - {}^tO \cdot D^2w(b) \cdot O$ to the 22-dimensional plane orthogonal to a and $O \cdot b$ has a simple structure spectrum which pemits to apply the first test of Lemma 4.2 to get a non-classical solution of a Hessian uniformly elliptic equation. Moreover, since $w_{\delta,ee}(a) = P_{ee}(a) - (1+\delta)P(a)$, for any $e \in S^{23}_1, e \perp a$, a more profound analysis of this restriction shows that the second part of Lemma is applicable as well for any $\delta \in]0, 1]$ which leads to singular solutions of such equations. In fact, since the multiplicity of some eigenvalues equals six one can descend to 21 dimensions keeping these properties [30]. Finally, a far more complicated analysis of the full Hessian D^2w_δ in 12 dimensions (which still has the factor $(x^3 - x + 2n)^2$ in its characteristic polynomial!) permits to apply the second part of Lemma 4.2 in 12 dimensions and its first part in 11 dimensions thus giving singular and non-classical solutions of a Hessian uniformly elliptic

equation in those dimensions [28, 30]. Moreover, one can formulate a test similar to the second part of Lemma 4.2 which garanties that w_δ is a solution to an Isaacs equation.

In this way we get the following:

1). For any δ, $1 \leq \delta < 2$ and any plane $H' \subset \mathbf{R}^{24}$, $\dim H' = 21$ the function

$$(P_{24}(x)/|x|^\delta)_{|H'}$$

is a viscosity solution to a uniformly elliptic Hessian (1.2) in the unit ball $B \subset \mathbf{R}^{21}$.

2). For any δ, $1 \leq \delta < 2$ the function

$$w_{12,\delta} = P_{12}(x)/|x|^\delta$$

is a viscosity solution to a uniformly elliptic Hessian equation (1.2) in the unit ball $B \subset \mathbf{R}^{12}$.

3). For any hyperplane $H \subset \mathbf{R}^{12}$ the function

$$(P_{12}(x)/|x|)_{|H}$$

is a viscosity solution to a uniformly elliptic Hessian equation (1.2) in the unit ball $B \subset \mathbf{R}^{11}$.

4). For any δ, $1 \leq \delta < 2$ the function

$$w_{12,\delta} = P_{12}(x)/|x|^\delta$$

is a viscosity solution to Isaacs equation (1.8) in the unit ball $B \subset \mathbf{R}^{12}$.

There remains a question on the minimal dimension n for which there exists a homogeneous order a, $1 < a \leq 2$, solutions of fully nonlinear uniformly elliptic equations. Lawson and Osserman's example [23] shows that Alexanrov's theorem does not hold in dimension 4. However, we expect that in dimension 4 there are still no such homogeneous solutions to fully nonlinear uniformly elliptic equations. One notes also that the constuction of Lawson and Osserman's example resembles strikingly that of w_{12}; it would be very intersting to clarify the reason underliyng that similarity.

5. Special Lagrangian Equation

In this section we study weak solutions a Hessian fully nonlinear second-order strictly, but not uniformly elliptic equations of the form (where $h \in \mathbf{R}$)

$$\mathbf{F}_h(D^2u) = \det(D^2u) - Tr(D^2u) + h\sigma_2(D^2u) - h = 0 \tag{5.1}$$

defined in a smooth-bordered domain of $\Omega \subset \mathbf{R}^3$, $\sigma_2(D^2u) = \lambda_1\lambda_2+\lambda_2\lambda_3+\lambda_1\lambda_3$ being the second symmetric function of the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of D^2u. This equation is equivalent to the Special Lagrangian potential equation [15]:

$$SLE_\theta : \qquad Im\{e^{-i\theta}\det(I + iD^2u)\} = 0 \tag{5.2}$$

for $h := -\tan(\theta)$ which can be re-written as

$$\mathbf{F}_\theta = \arctan\lambda_1 + \arctan\lambda_2 + \arctan\lambda_3 - \theta = 0.$$

The set

$$\{A \in Sym^2(\mathbf{R}^3) : \mathbf{F}_h(A) = 0\} \subset Sym^2(\mathbf{R}^3)$$

has three connected components, $C_i, i = 1, 2, 3$ which correspond to the values $\theta_1 = -\arctan(h) - \pi, \theta_2 = -\arctan(h), \theta_3 = -\arctan(h) + \pi$.

We study the Dirichlet problem

$$\begin{cases} \mathbf{F}_h(D^2u) = 0 & \text{in } \Omega \\ u = \varphi & \text{on } \partial\Omega \end{cases} \tag{5.3}$$

where $\Omega \subset \mathbf{R}^n$ is a bounded domain with smooth boundary $\partial\Omega$ and φ is a continuous function on $\partial\Omega$.

For $\theta_1 = -\arctan(h) - \pi$ and $\theta_3 = -\arctan(h) + \pi$ the operator $\mathbf{F}_\theta$ is concave or convex, and the Dirichlet problem in these cases was treated in [10]; smooth solutions are established there for smooth boundary data on appropriately convex domains.

The middle branch $C_2, \theta_2 = -\arctan(h)$ is never convex (neither concave), and the classical solvability of the Dirichlet problem remained open.

In the case of uniformly elliptic equations a theory of weak (viscosity) solutions for the Dirichlet problem gives the uniqueness of such solutions, see [9]. One can define viscosity solutions for non-uniformly elliptic equations (such as SLE_θ) as well, but in this case the uniqueness of viscosity solutions known to experts in the field is not given explicitly in the literature, so we use a new very interesting approach to degenerate elliptic equations suggested recently by Harvey and Lawson [16]. They introduced a new notion of a weak solution for the Dirichlet problem for such equations and proved the existence, the continuity and the uniqueness of these solutions.

We are going to explain here why the classical solvability for Special Lagrangian Equations *does not* hold. More precisely, for any $\theta \in]-\pi/2, \pi/2[$ there exist a small ball $B \subset \mathbf{R}^3$ and an analytic function φ on ∂B for which the unique Harvey-Lawson solution u_θ of the Dirichlet problem satisfies:

(i) $u_\theta \in C^{1,1/3}$;

(ii) $u_\theta \notin C^{1,\delta}$ for $\forall\delta > 1/3$.

Our construction use the Legendre transform for solutions of $\mathbf{F}_{\frac{1}{h}}(D^2u) = 0$ which gives solutions of $\mathbf{F}_h(D^2u) = 0$; in particular, for $h = 0$ it transforms the solutions of $\sigma_2(D^2u) = 1$ into solutions of $\det(D^2u) = Tr(D^2u)$.

The costruction is based on the following result which can be verified by a direct calculation along with the Cauchy-Kowalevskaya theorem :

Lemma 5.1. *There exists a ball $B = B(0, \varepsilon)$ centered at the origin s.t. the equation*

$$\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 = \sigma_2(D^2u) = 1$$

has an analytic solution u_0 in B verifying

(i)

$$u_0 = -\frac{y^4}{3} + 5y^2z^2 - x^4 + 7x^2z^2 - \frac{z^4}{3} + 2y^2z - 2zx^2 + \frac{y^2}{2} + \frac{x^2}{2} + O(r^5)$$

(ii)

$$\lambda_1 = 1 + O(r), \;\; \lambda_2 = 1 + O(r), \;\; \lambda_3 = -\frac{x^2}{2} - \frac{3y^2}{2} - z^2 + O(r^3).$$

This lemma corresponds to $h = 0$; but the same statement is true for all h with appropriate functions u_h similar to u_0. It is possible to show that ∇u_h is bijective on a small ball and its Legendre transform verifies the conditions (i) and (ii).

Recall that the Special Lagrangian equation is equivalent to the condition that the graph of ∇u is minimal [15] . However, for our solutions u_θ the corresponding graph is smooth, and the singularity of solutions correspond to a singularity of the projection of this graph onto the domain of definition. It would be interesting to know whether it is always the case. Recall that the subject of singular special Lagrange submanifolds became very popular due to its possible connection to the mirror phenomena [37].

One can also ask whether there exists a non-classical solution of the Special Lagrangian equation similar to those studied higher for uniformly elliptic equations, i.e. homogenous of order two. It follows from the main result of [18] that such solution does not exist in any dimension.

References

[1] J.F. Adams, *Lectures on Exceptional Lie Groups*, eds. Z. Mahmud and M. Mimira, University of Chicago Press, Chicago, 1996.

[2] A.D. Alexandroff, *Sur les théorèmes d'unicite pour les surfaces fermées*, Dokl. Acad. Nauk 22 (1939), 99–102.

[3] J. Baez, *Octonions*, Bull. Amer. Math. Soc. 39 (2001), 145–205.

[4] G.Barles, F. Da Lio, *Local $C^{0,\alpha}$ estimates for viscosity solutions of Neumann-type boundary value problems,* J. Diff. Eq. 225 (2006) 202–224.

[5] R. Bass, E. Pardoux, *Uniqueness for the diffusion with piecewise constant coefficients,* Probability Theory Related Fields 76 (1987), 557–572.

[6] L. Caffarelli, X. Cabre, *Fully Nonlinear Elliptic Equations*, Amer. Math. Soc., Providence, R.I., 1995.

[7] L. Caffarelli, M.G. Crandall, M.Kocan, A. Swiech, *On viscosity solutions of fully nonlinear elliptic equations with measurable ingredients,* Comm. Pure Appl. Math. 49 (1996), 365–397.

[8] M.C. Cerutti, L. Escauriaza, E.B.Fabes, *Uniqueness in the Dirichlet problem for some elliptic operators with discontinuous coefficients,* Ann. Math. Pura Appl. 163 (1996), 161–180.

[9] M.G. Crandall, H. Ishii, P-L. Lions, *User's guide to viscosity solutions of second order partial differential equations,* Bull. Amer. Math. Soc. (N.S.), 27(1) (1992), 1–67.

[10] L. Caffarelli, L. Nirenberg, J. Spruck, *The Dirichlet problem for nonlinear second order elliptic equations III. Functions of the eigenvalues of the Hessian,* Acta Math. 155 (1985), no. 3–4, 261–301.

[11] L. C. Evans, *Classical solutions of fully nonlinear, convex, second-order elliptic equations* , Comm. Pure Appl. Math. 35 (1982), 333–363.

[12] W.H. Fleming, P.E. Souganidis *On the existence of value functions of two-player, zero-sum stochastic differential games,* Indiana Univ. Math. J., 33 (1989), 293–314.

[13] D. Gilbarg, N. Trudinger, *Elliptic Partial Differential Equations of Second Order,* 2nd ed., Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 1983.

[14] Q. Han, N. Nadirashvili, Y. Yuan, *Linearity of homogeneous order-one solutions to elliptic equations in dimension three,* Comm. Pure Appl. Math. 56 (2003), 425–432.

[15] R. Harvey, H. B. Lawson Jr., *Calibrated geometries,* Acta Math. 148 (1982), 47–157.

[16] F. R. Harvey, H. B. Lawson Jr., *Dirichlet duality and the nonlinear Dirichlet problem*, Comm. Pure Appl. Math. 62 (2009), no 3, 396–443.

[17] R. Jensen, *Uniformly elliptic PDE's with measurable coefficients,* J. Fourier Anal. Appl. 2 (1996), 237–259.

[18] J. Jost., Y.-L. Xin, *A Bernstein theorem for special Lagrangian graphs,* Calc. Var. Part. Diff. Eq. 15 (2002), 299–312.

[19] N.V. Krylov, *Controlled Diffusion Process,* Springer Verlag 1980.

[20] N.V. Krylov, *Nonlinear Elliptic and Parabolic Equations of Second Order*, Reidel, 1987.

[21] N.V. Krylov, *On weak uniqueness for some diffusions with discontinuous coefficients,* Stochastic Proc. Appl. 113 (2004), 37–64.

[22] C.E. Kenig, N. Nadirashvili, *On optimal estimates for some oblique derivative problems,* J. Funct. Anal. 187 (2001), 70–93.

[23] H. B. Lawson and R. Osserman, *Non-existence, non-uniqueness and irregularity of solution to the minimal surface system,* Acta Math., 139 (1977), 1–17.

[24] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. I.The dynamic programming principle and application,* Comm. Part. Diff. Eq. 8 (1983), 1101–1174.

[25] N. Nadirashvili, *Nonuniqueness in the martingale problem and the Dirichlet problem for uniformly elliptic operators* , Ann. Sc. Norm. Sup. Pisa, Ser. 4, 24 (1997), 537–550.

[26] N. Nadirashvili, S. Vlăduţ, *Nonclassical solutions of fully nonlinear elliptic equations,* Geom. Func. An. 17 (2007), 1283–1296.

[27] N. Nadirashvili, S. Vlăduţ, *Singular solutions to fully nonlinear elliptic equations,* J. Math. Pures Appl. 89 (2008), 107–113.

[28] N. Nadirashvili, S. Vlăduţ, *On Hessian fully nonlinear elliptic equations*, arXiv:0805.2694 [math.AP], submitted.

[29] N. Nadirashvili, S. Vlăduţ, *Singular solution to special lagrangian equations,* to appear in: Annales de l'Institut Henri Poincare (C) Non Linear Analysis .

[30] N. Nadirashvili, S. Vlăduţ, *Nonclassical Solutions of Fully Nonlinear Elliptic Equations II. Hessian Equations and Octonions*, arXiv:0912.3126, submitted.

[31] N. Nadirashvili, S. Vlăduţ, *On Axially Symmetric Solutions of Fully Nonlinear Elliptic Equations*, arXiv:1003.0032, submitted.

[32] N. Nadirashvili, Y. Yuan, *Homogeneous solutions to fully nonlinear elliptic equation*, Proc. AMS, 134:6 (2006), 1647–1649.

[33] L. Nirenberg, *On nonlinear elliptic partial differential equations and Hölder continuity,* Comm. Pure Appl. Math. 6 (1953), 103–156.

[34] M.V. Safonov, *Nonuniqueness for second-order elliptic equations with measurable coefficients,* SIAM J. Math. Anal. 30 (1999), 879–895.

[35] M.V. Safonov, *On weak uniqueness for some elliptic equations,* Comm. Part. Diff. Eq. 19 (1994), 943–957.

[36] D.V. Stroock, S.R.S. Varadhan, *Multidimensional Diffusion Processes,* 2nd ed., Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 1997.

[37] A. Strominger, S.-T. Yau, E. Zaslow, *Mirror Symmetry is T-duality,* Nucl. Phys. B 479 (1996), 243–259.

[38] N. Trudinger, *Weak solutions of Hessian equations,* Comm. Partial Differential Equations 22 (1997), no. 7–8, 1251–1261.

[39] N. Trudinger, *On the Dirichlet problem for Hessian equations,* Acta Math. 175 (1995), no. 2, 151–164.

[40] G. Weyl, *Das asymptotische Verteilungsgezets des Eigenwerte lineare partieller Differentialgleichungen,* Math. Ann. 71 (1912), no. 2, 441–479.

Section 12

Mathematical Physics

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Topological Field Theory, Higher Categories, and Their Applications

Anton Kapustin*

Abstract

It has been common wisdom among mathematicians that Extended Topological Field Theory in dimensions higher than two is naturally formulated in terms of n-categories with $n > 1$. Recently the physical meaning of these higher categorical structures has been recognized and concrete examples of Extended TFTs have been constructed. Some of these examples, like the Rozansky-Witten model, are of geometric nature, while others are related to representation theory. I outline two applications of higher-dimensional TFTs. One is related to the problem of classifying monoidal deformations of the derived category of coherent sheaves, and the other one is geometric Langlands duality.

Mathematics Subject Classification (2010). Primary 57R56; Secondary 81T45, 18D05, 14D24, 14F05

Keywords. Topological field theory, 2-categories, monoidal categories, derived category of coherent sheaves, geometric Langlands duality

1. Introduction

The notion of functional integral[1] plays a central role in quantum field theory, but it has defied attempts at a rigorous mathematical formulation, except in some special cases (typically in space-time dimension 2). Topological Field Theory (TFT) provides a useful playground for studying properties of the functional integral in a simplified setting and has been the subject of many works since the pioneering papers by E. Witten [35, 36, 37]. The physical definition

*California Institute of Technology, Pasadena, CA 91125, United States.
E-mail: kapustin@theory.caltech.edu.

[1]The term "functional integral" is synonymous with "path-integral", but is more descriptive, since in Quantum Field Theory one integrates over a space of functions of several variables rather than over a space of paths.

of the functional integral uses an ill-defined measure on the space of field configurations, but one can use the usual mathematical ploy and try to axiomatize properties of the functional integral without making a direct reference to this measure. The first attempt at such an axiomatization was made by M. Atiyah [2]. Atiyah defines a TFT in n dimensions as a functor F from a certain geometrically defined category Bord_n to the category of complex vector spaces Vect (or to the category of $\mathbb{Z}_2$-graded complex vector spaces $\mathrm{Vect}_{\mathbb{Z}_2}$). The category Bord_n has as its objects compact oriented $(n-1)$-manifolds without boundary and has as its morphisms oriented bordisms between such manifolds. F is supposed to be invariant with respect to diffeomorphisms. The disjoint union gives the category Bord_n a symmetric monoidal structure whose identity object is the empty $(n-1)$-manifold. The category Vect also has a natural symmetric monoidal structure given by the tensor product; the identity object is the field $\mathbb{C}$. The functor F is required to be monoidal; in particular, it sends the disjoint union of two $(n-1)$-manifolds M_1 and M_2 to the tensor product $V(M_1) \otimes V(M_2)$, and it sends the empty manifold to $\mathbb{C}$.

Since we can regard a closed oriented n-manifold N as a bordism between $\emptyset$ and $\emptyset$, the functor F sends any such N to a linear map from $\mathbb{C}$ to $\mathbb{C}$, i.e. a complex number $\mathsf{F}(N)$. This number is called the partition function of the TFT on the manifolds N. So we see that an n-dimensional TFT assigns a number to a closed oriented n-manifold and a vector space to a closed oriented $(n-1)$-manifold.

It is natural to ask if an n-dimensional TFT assigns anything to closed oriented manifolds of lower dimensions. An obvious guess is that it assigns a $\mathbb{C}$-linear category to an $(n-2)$-manifold, a $\mathbb{C}$-linear 2-category to an $(n-3)$-manifold, etc. The resulting gadget is usually called an *Extended Topological Field Theory*.[2] Extending the TFT functor to lower-dimensional manifolds is natural if we consider gluing closed manifolds out of manifolds with boundaries. For example, given two oriented $(n-1)$-dimensional manifolds N_1 and N_2 and an orientation-reversing diffeomorphism $\partial N_1 \to \partial N_2$, we can glue N_1 and N_2 along their common boundary and get a closed oriented $(n-1)$-manifold N_{12}. An Extended TFT in n dimensions assigns to ∂N_1 a $\mathbb{C}$-linear category $\mathsf{F}(\partial N_1)$ and assigns to N_1 and $\bar{N}_2$ (the orientation-reversal of N_2) objects $\mathsf{F}(N_1)$ and $\mathsf{F}(\bar{N}_2)$ of this category. The fact that N_{12} can be glued from N_1 and N_2 means that the vector space $\mathsf{F}(N_{12})$ is the space of morphisms from the object $\mathsf{F}(N_1)$ to the object $\mathsf{F}(\bar{N}_2)$.

While the relevance of higher categories for TFT in higher dimensions has been recognized by experts for some time [8, 9, 22, 3], an axiomatic definition of an Extended TFT has not been formulated for an obvious reason: the lack

[2]It is likely that the language of (∞, n)-categories whose theory is being developed by J. Lurie [23] is even better suited for TFT applications [24, 11]. Its physical significance remains unclear at the time of writiing.

of a universally accepted definition of an n-category for $n > 2$.[3] This technical obstacle was compounded by a lack of understanding of the physical meaning of higher categories.

The correct definition of a weak n-category being non-obvious, one is forced to go back to the physical roots of the subject. We will first discuss two-dimensional TFTs which have been studied extensively because of their connection with Mirror Symmetry and explain why boundary conditions in a 2d TFT form a category. This observation is due to M. Douglas [7]. Then we will move on to three dimensions and explain why boundary conditions in a 3d TFT form a 2-category. Applying these observations to n-dimensional TFT we will be able to see from a more physical viewpoint why n-dimensional TFT assigns a $\mathbb{C}$-linear $(k-1)$-category to a compact oriented $(n-k)$-manifold. Then we will describe two examples of TFTs in three and four dimensions and their applications to two different mathematical problems: the classification of monoidal deformations of the derived category of coherent sheaves and the Geometric Langlands Program.

2. Extended Topological Field Theory from a physical viewpoint

2.1. Extended TFT in two dimensions. Consider a 2d TFT on a compact oriented 2-manifold Σ with a nonempty boundary. It turns out necessary to impose some conditions on the values of the fields on $\partial\Sigma$ for the functional integral to be well-defined on the physical level of rigor. Roughly speaking, these conditions must define a Lagrangian submanifold in the space of boundary values of the fields, where the symplectic form arises from the boundary terms in the variation of the action. On the classical level, boundary conditions are needed to make the initial-value problem for the classical equations of motion well-posed and to ensure the existence of a symplectic form on the space of solutions.

Boundary conditions in a 2d TFT (also known as branes) form a $\mathbb{C}$-linear category. This is the category which the 2d TFT assigns to a point. Morphisms in this category are *boundary-changing local operators*. To explain informally what this means, suppose Σ is a half-plane $\{(x, y) \in \mathbb{R}^2 | x \geq 0\}$, and one imposed a boundary condition A on the half-line $\{(0, y) | y < 0\}$ and a boundary condition B on the half-line $\{(0, y) | y > 0\}$ (see Fig. 1). At the special point $(0, 0)$ additional data are needed to specify the functional integral uniquely. These data define a boundary-changing point operator $\mathcal{O}_{AB}$ between A and B.

[3] We remind that one distinguishes strict and weak n-categories. While the former are easily defined, they almost never occur in practice; to define Extended TFTs one needs weak n-categories.

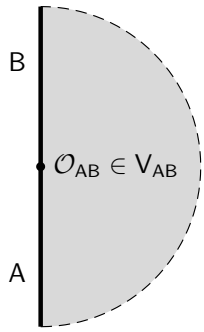


Figure 1. Morphisms in the category of boundary conditions are boundary-changing point operators.

It is a basic physical principle that the set of boundary-changing point operators has the structure of a (graded) vector space. To see why, let us introduce polar coordinates (r, ϕ), $r \in \mathbb{R}_+$, $\phi \in [-\pi/2, \pi/2]$, so that the origin $(0, 0)$ is given by $r = 0$. To avoid dealing with divergences ubiquitous in quantum field theory one may cut out a small half-disc $r < \epsilon$ for some $\epsilon > 0$ and replace the boundary-changing point operator by a suitable boundary condition on a semicircle $r = \epsilon$ (Fig. 2). (Unlike the boundary conditions A and B the boundary condition corresponding to a boundary-changing point operator is nonlocal, in general, in the sense that it does not merely constrain the values of the fields and a finite number of their derivatives along the boundary but may involve constraints on the Fourier components of the restrictions of the fields to the boundary.) The key remark is that the half-plane with a half-disc removed is diffeomorphic to a product of a half-line parameterized by r, $r \geq \epsilon$, and the interval $[-\pi/2, \pi/2]$ parameterized by ϕ. We now reinterpret r as the time coordinate and ϕ as the spatial coordinate. On the spatial boundaries $\phi = \mp\pi/2$ we have boundary conditions A and B, while the boundary condition at $r = \epsilon$ is now regarded as an initial condition. Initial states in any quantum theory form a vector space (in fact, a Hilbert space). We conclude that boundary-changing point operators between boundary conditions A and B form a vector space V_{AB}.[4]

Next we consider a situation where Σ is a half-plane, but its boundary is divided into three pieces: a half-line $y < 0$, an interval $0 < y < a$, and a half-line $y > a$. We impose boundary conditions $\mathsf{A}, \mathsf{B}, \mathsf{C}$ on the three pieces respectively, so we need two boundary-changing point operators which are elements of vector spaces V_{AB} and V_{BC} (Fig. 3). In a TFT the limit $a \to 0$ always exists, and one should be able to interpret these two boundary-changing point operators as a single boundary-changing point operator between A and C, i.e. an element of

[4]Note that orientation is important here, so V_{AB} is not the same as V_{BA}.

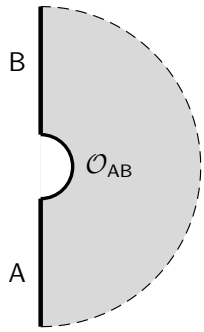


Figure 2. A boundary-changing point operator is equivalent to a (possibly nonlocal) boundary condition.

the vector space V_{AC}. This gives a 'fusion product"

$$\mathsf{V}_{\mathsf{AB}} \times \mathsf{V}_{\mathsf{BC}} \to \mathsf{V}_{\mathsf{AC}}$$

One may further argue that this product is bilinear and associative in an obvious sense. Note also that for any A the vector space V_{AA} has a special element: the boundary-changing point operator which is trivial. This element serves as a unit in the algebra V_{AA}. Altogether we obtain a category whose objects are branes, whose morphisms are elements of vector spaces V_{AB}, and with the composition of morphisms defined by means of the fusion product.

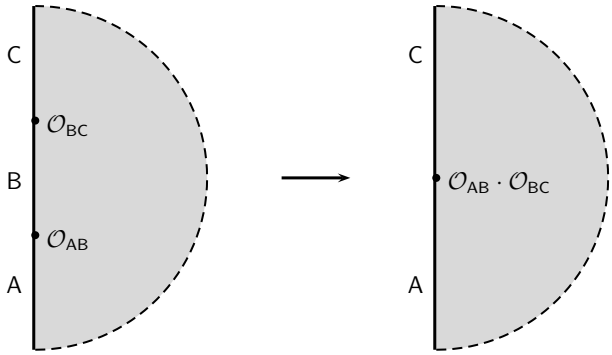


Figure 3. Composition of morphisms corresponds to fusing boundary-changing point operators. Fusion product is denoted by a dot.

Axioms of 2d TFT as usually formulated further imply that the resulting category is self-dual, in the sense that the spaces V_{AB} and V_{BA} are naturally dual, but we will not emphasize this aspect, since in some cases with infinite-dimensional spaces V_{AB} this requirement is not satisfied.

The last remark we want to make about Extended TFTs in two dimensions is that instead of closed oriented manifolds one may consider compact oriented manifolds with a nonempty boundary. The connected components of the boundary should be labeled by objects of the category of branes. We will call this labeling a decoration of a manifold. Extended TFT assigns a complex number to a decorated 2-manifold (the value of the functional integral with the corresponding boundary conditions). It assigns a vector space to a decorated 1-manifold. The only connected decorated 1-manifold is an interval, whose decoration consists of an ordered pair (A,B) of branes. The vector space assigned to such a pair is the space of boundary-changing local operators V_{AB} introduced above. A good way to think about this rule is the following: if we consider our 2d TFT on a 2-manifold of the form $\mathbb{R}\times[0,1]$, where the two connected components of the boundary are labeled by A and B, then we may regard it as a 1d TFT on $\mathbb{R}$ (this is called Kaluza-Klein reduction). A 1d TFT is simply a quantum mechanical system, and its Hilbert space of states is what we assign to the interval $[0,1]$.

2.2. Extended TFT in three dimensions. Three-dimensional TFT is supposed to assigns a $\mathbb{C}$-linear category to a closed oriented 1-manifold and a $\mathbb{C}$-linear 2-category to a point. Let us first explain the physical meaning of the former. Consider a 3d TFT on a manifold of the form $S^1\times\Sigma$, where Σ is an oriented 2-manifold which may be noncompact or have a nonempty boundary. Another basic physical principle (Kaluza-Klein reduction) is that in such a case one can describe the physics of the compactified theory by an *effective 2d TFT* on Σ. By definition, the category assigned to a circle is the category of branes in this effective 2d TFT.

The 2-category assigned to a point is the 2-category of boundary conditions in the 3d TFT. To explain where the 2-category structure comes from, consider a 3d TFT on an oriented 3-manifold W with a nonempty boundary, imagine that a connected component of ∂W is subdivided by closed curves into domains, and that one imposed unrelated boundary conditions on different domains. Each domain is thus labeled by an element of the set of boundary conditions. A closed curve separating the domains labeled by boundary conditions $\mathbb{X}$ and $\mathbb{Y}$ is itself labeled by an element of a set $\mathsf{W}_{\mathbb{X}\mathbb{Y}}$ which determines how fields behave in the neighborhood of the closed curve. Elements of the set $\mathsf{W}_{\mathbb{X}\mathbb{Y}}$ will be called boundary-changing line operators from $\mathbb{X}$ to $\mathbb{Y}$. Boundary-changing line operators may be fused together (Fig. 4) which gives rise to a fusion product

$$\mathsf{W}_{\mathbb{X}\mathbb{Y}}\times\mathsf{W}_{\mathbb{Y}\mathbb{Z}}\to\mathsf{W}_{\mathbb{X}\mathbb{Z}},\quad(\mathsf{A},\mathsf{B})\mapsto\mathsf{A}\otimes\mathsf{B},\ \forall\mathsf{A}\in\mathsf{W}_{\mathbb{X}\mathbb{Y}},\forall\mathsf{B}\in\mathsf{W}_{\mathbb{Y}\mathbb{Z}}. \tag{1}$$

In every set $\mathsf{W}_{\mathbb{X}\mathbb{X}}$ there is a special element, the trivial boundary-changing line operator, which is an identity element with respect to the fusion product. The associativity of the fusion product is more difficult to formulate because there are boundary-changing line operators which are physically equivalent, but not equal. From the mathematical point of view, the difficulty can be explained

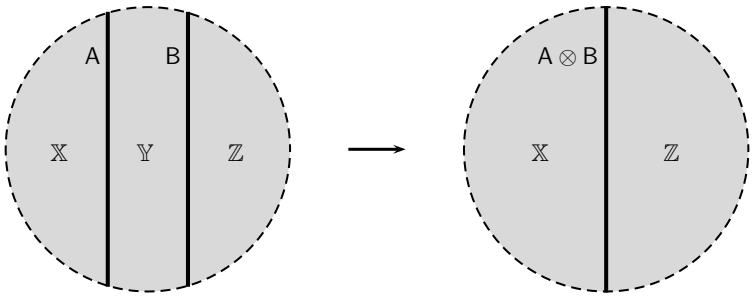


Figure 4. Boundary-changing line operator A from a boundary condition $\mathbb{X}$ to a boundary condition $\mathbb{Y}$ and boundary-changing line operator B from $\mathbb{Y}$ to a $\mathbb{Z}$ can be fused to produce a boundary-changing line operator from $\mathbb{X}$ to $\mathbb{Z}$. This operation is denoted $\otimes$.

by saying that the set $\mathsf{W}_{\mathbb{X}\mathbb{Y}}$ is actually a category, and it is not natural to talk about equality of objects in a category. Morphisms in this category are point operators inserted at points on the closed curve separating domains $\mathbb{X}$ and $\mathbb{Y}$ (Fig. 5). The insertion points of point operators divide the closed curve into segments, and each segment is labeled by an element of the set $\mathsf{W}_{\mathbb{X}\mathbb{Y}}$.

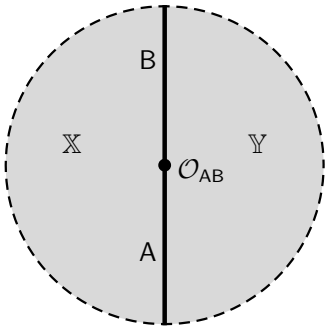


Figure 5. Boundary-changing line operators between boundary conditions $\mathbb{X}$ and $\mathbb{Y}$ are objects of a category $\mathsf{W}_{\mathbb{X}\mathbb{Y}}$. A morphism $\mathcal{O}_{\mathsf{AB}}$ from an object A to an object B is a point operator inserted at the junction of A and B.

Let A and B be boundary-changing line operators between $\mathbb{X}$ and $\mathbb{Y}$; we will denote by V_{AB} the set of point operators which can be inserted at the junction of segments labeled by A and B. The same reasoning as in the case of boundary-changing point operators tells us that V_{AB} is a vector space. Point operators sitting on the same closed curve can be fused (Fig. 6), which gives rise to a product

$$\mathsf{V}_{\mathsf{AB}} \times \mathsf{V}_{\mathsf{BC}} \to \mathsf{V}_{\mathsf{AC}},$$

which is bilinear and associative in an obvious sense. The category of boundary-changing line operators between $\mathbb{X}$ and $\mathbb{Y}$ has $\mathsf{W}_{\mathbb{X}\mathbb{Y}}$ as its set of objects and the sets V_{AB} as the sets of morphisms. Let us denote this category $\mathcal{C}_{\mathbb{X}\mathbb{Y}}$.

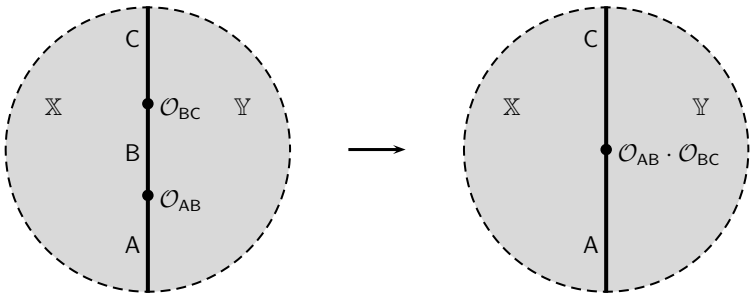


Figure 6. Composition of morphisms in the category $\mathsf{W}_{\mathbb{X}\mathbb{Y}}$ arises from the fusion of point operators sitting at the junctions of boundary-changing line operators.

The proper formulation of associativity of the fusion product (1) says that that two triple products of three boundary-changing line operators differing by a placement of parentheses are isomorphic:

$$(\mathsf{A} \otimes \mathsf{B}) \otimes \mathsf{C} \simeq \mathsf{A} \otimes (\mathsf{B} \otimes \mathsf{C}), \quad \forall \mathsf{A} \in \mathsf{W}_{\mathbb{X}\mathbb{Y}}, \forall \mathsf{B} \in \mathsf{W}_{\mathbb{Y}\mathbb{Z}}, \forall \mathsf{C} \in \mathsf{W}_{\mathbb{Z}\mathbb{T}}.$$

The isomorphism must be specified and must satisfy the so-called pentagon identity [25]. The above discussion can be summarized by saying that boundary conditions in a 3d TFT form a 2-category, whose sets of 1-morphisms are sets $\mathsf{W}_{\mathbb{X}\mathbb{Y}}$ and whose sets of 2-morphisms are vector spaces V_{AB}.

Kaluza-Klein reduction enables us to think about the category $\mathcal{C}_{\mathbb{X}\mathbb{Y}}$ in two-dimensional terms. Consider a 3d TFT on a 3-manifold of the form $\Sigma \times [0,1]$ where Σ is an oriented but not necessarily closed 2-manifold. On the boundaries $\Sigma \times \{0\}$ and $\Sigma \times \{1\}$ we impose boundary conditions $\mathbb{X}$ and $\mathbb{Y}$ respectively. Kaluza-Klein reduction tells us that one can describe the physics of this 3d TFT by an effective 2d TFT on Σ which depends on $\mathbb{X}$ and $\mathbb{Y}$. We claim that the category $\mathcal{C}_{\mathbb{X}\mathbb{Y}}$ is the category of branes for this effective 2d TFT. Indeed, consider a 3d TFT on the half-space $\{(x,y,z)|x \geq 0\}$, where we imposed the boundary condition $\mathbb{X}$ on the half-plane $\{(0,y,z)|y < 0\}$ and the boundary condition $\mathbb{Y}$ on the half-plane $\{(0,y,z)|y > 0\}$. At the line given by $x = y = 0$ we insert some boundary-changing line operator $\mathsf{A} \in \mathsf{W}_{\mathbb{X}\mathbb{Y}}$. To regularize the problem we need to cut out a solid half-cylinder $x^2 + y^2 < \epsilon^2$ for some $\epsilon > 0$ and replace A with a suitable boundary condition on the part of the boundary given by $x^2 + y^2 = \epsilon^2$ (Fig. 7). Now we note that the half-space with a solid half-cylinder removed is diffeomorphic to $\mathbb{R}_+ \times \mathbb{R} \times [-\pi/2, \pi/2]$, where $\mathbb{R}_+$ is parameterized by the radial coordinate on the (x,y) plane, $\mathbb{R}$ is parameterized by z, and the interval is parameterized by the angular coordinate on the (x,y) plane. Thus we may interpret the boundary condition at $x^2 + y^2 = \epsilon^2$ representing a boundary-

changing line operator as a boundary condition in the effective 2d TFT on the half-space $\mathbb{R}_+ \times \mathbb{R}$.

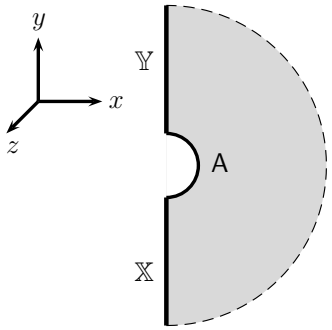


Figure 7. A boundary-changing line operator is equivalent to a boundary condition on a half-cylinder $x^2 + y^2 = \epsilon^2, x > 0$. This boundary condition is local in the z direction but may be nonlocal in the angular direction in the xy plane. It can be interpreted as a local boundary condition in a 2d TFT which is obtained by reducing the 3d TFT on an interval.

For any object $\mathbb{X}$ of a 2-category the endomorphism category $\mathcal{C}_{\mathbb{X}\mathbb{X}}$ has a monoidal structure, i.e. an associative but not necessarily commutative tensor product. This monoidal structure is not natural from the 2d viewpoint (there is no physically reasonable way to define tensor product of branes in a general 2d TFT). Turning this around, if there is a mathematically natural monoidal structure on a category of branes in a 2d TFT, it is likely that this 2d TFT arises as a Kaluza-Klein reduction of a 3d TFT on an interval, and its category of branes can be interpreted as the category $\mathcal{C}_{\mathbb{X}\mathbb{X}}$ for some boundary condition $\mathbb{X}$ in this 3d TFT. We will see an example of this below.

As in 2d TFT, we may consider decorated manifolds, i.e. compact oriented manifolds with a nonempty boundary whose connected components are labeled by elements of the set of boundary conditions. Extended TFT in three dimensions assigns a number to a decorated 3-manifold (the value of the functional integral with given boundary conditions), a vector space to a decorated 2-manifold (the space of states of the effective 1d TFT obtained by Kaluza-Klein reduction on this 2-manifold), and a category to a decorated 1-manifold. The only connected decorated 1-manifold is an interval $[0, 1]$. If its endpoints are labeled by boundary conditions $\mathbb{X}$ and $\mathbb{Y}$, the corresponding category is the category of boundary-changing line operators $\mathcal{C}_{\mathbb{X}\mathbb{Y}}$.

2.3. Extended TFT in n dimensions. Continuing in the same fashion we conclude that boundary conditions in an n-dimensional TFT form an $(n-1)$-category. By analyzing more precisely the physical notion of a boundary condition one should be able to arrive at a physically-motivated definition of a weak n-category for all $n > 0$. We will not try to do it here.

We can now see why n-dimensional TFT assigns a k-category to a closed oriented $(n-k-1)$-manifold M. Consider an n-dimensional TFT on a manifold of the form $M \times N$, where N is an oriented but not necessarily closed $(k+1)$-manifold. The Kaluza-Klein reduction principle tells us that we can describe the physics by an effective $(k+1)$-dimensional TFT on N. The k-category assigned to M is the k-category of boundary conditions for this effective $(k+1)$-dimensional TFT.

If M is the $(n-k-1)$-dimensional sphere, the corresponding k-category has an alternative interpretation: it is the category of defects of dimension k. To explain what a defect is, we might imagine that our TFT describes a particular macroscopic quantum state of a system of atoms in space-time of dimension n. It may happen that along some oriented submanifold L of dimension k the atoms are in a different state than elsewhere (or perhaps there is an altogether different kind of atoms inserted along this submanifold). In such a case one says that there is a defect of dimension k inserted at L. Zero-dimensional defects are also known as local operators, one-dimensional defects are known as line operators, two-dimensional defects are known as surface operators.

We claim that defects of dimension k form a k-category, and that this category is nothing but the k-category assigned to S^{n-k-1}. To see this, suppose that L is a k-dimensional plane in $\mathbb{R}^n$. We introduce "polar" coordinates in $\mathbb{R}^n$ such that $\mathbb{R}^n \backslash L$ is identified with $\mathbb{R}^k \times S^{n-k-1} \times \mathbb{R}_+$, and L is given by $r = 0$, where r is the coordinate on $\mathbb{R}_+$. To regularize the problem we usually need to cut out a small tubular neighborhood of L given by $r < \epsilon$ for some $\epsilon > 0$ and replace the defect by a suitable boundary condition at $r = \epsilon$. Since our n-manifold has a factor S^{n-k-1}, we may regard this boundary condition as a boundary condition in an effective $(k+1)$-dimensional TFT which is obtained by Kaluza-Klein reduction on S^{n-k-1}. Thus defects of dimension k can be regarded as objects of the k-category assigned to S^{n-k-1}. We will denote it $\mathcal{D}_k$.

Let us note a few special cases. Local operators (i.e. defects of dimension 0) are elements of the vector space assigned to S^{n-1}. This is usually called the state-operator correspondence. Line operators are objects of a $\mathbb{C}$-linear category, while surface operators are objects of a $\mathbb{C}$-linear 2-category.

Two defects of dimension k can be fused, which gives rise to a monoidal structure on $\mathcal{D}_k$. Another way to understand the origin of this monoidal structure is to note that a solid ball of dimension $n-k$ with two smaller balls removed gives a canonical bordims from $S^{n-k-1} \sqcup S^{n-k-1}$ and S^{n-k-1} (Fig. 8). The Extended TFT assigns to this bordism a k-functor from $\mathcal{D}_k \times \mathcal{D}_k$ to $\mathcal{D}_k$. This functor is associative (in some nontrivial sense). If $n-k$ is greater than 2 it is also commutative, while for $n-k=2$ it is only braided. Thus monoidal, braided monoidal and symmetric monoidal k-categories also naturally arise from Extended TFT. We note that M. Kapranov and V. Voevodsky proposed definitions for monoidal and braided monoidal 2-categories in [16]. It

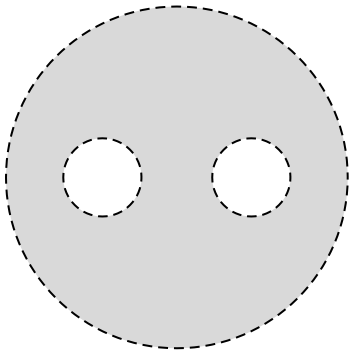


Figure 8. A canonical bordism from $S^{n-k-1} \sqcup S^{n-k-1}$ to S^{n-k-1}, here shown for $n-k=2$.

is not clear if these definition agree with the definitions which are natural from the point of view of Extended TFT.

3. The Rozansky-Witten model

3.1. Definition and basic properties. The Rozansky-Witten model is a 3d TFT whose definition and basic properties have been described in [31]. It is a 3d sigma-model, i.e. its only bosonic field is a map $\phi : M \to X$, where M is an oriented 3-manifold, and X is a complex manifold equipped with a holomorphic symplectic form. There are also fermionic fields: $\eta \in \Gamma(\phi^* T^{0.1} X)$ and $\rho \in \Gamma(\phi^* T^{1,0} X \otimes T^* M)$. The partition function of the RW model is defined as a functional integral over the infinite-dimensional supermanifold $\mathcal{M}$ with local coordinates (ϕ, η, ρ). The measure is proportional to $\exp(-S(\phi, \eta, \rho))$, where S is the action functional. Its explicit form is given in [31] but we will not need it here. The most important property of this measure is that it is invariant with respect to a certain odd vector field Q on $\mathcal{M}$ satisfying

$$\{Q, Q\} = 0,$$

where braces denote the super-Lie bracket. As a result the partition function of the theory is unchanged if one adds to S a function of the form $Q(f)$, where f is any sufficiently well-behaved function on $\mathcal{M}$. The Rozansky-Witten model is topological because its action can be written as

$$S = Q(V) + S_0,$$

where S_0 is independent of the metric on M and V is some metric-dependent odd function.

The Rozansky-Witten model has a formal similarity to the better-known Chern-Simons gauge theory [32, 37]. For example, the non-Q-exact part of the action reads

$$S_0 = \int_M \left(\Omega(\rho, D\rho) + \frac{1}{3}\Omega(\rho, R(\rho, \rho, \eta)) \right),$$

where Ω is the holomorphic symplectic form on X, D is the covariant differential on $\phi^* T^{1,0}X \otimes T^*M$ with respect to a pull-back of a connection on $T^{1,0}X$, and $R \in \mathrm{Hom}(T^{1,0}X \otimes T^{1,0}X \otimes T^{0,1}X, T^{1,0}X)$ is the curvature of this connection. Compare this to the action of the Chern-Simons gauge theory:

$$S_{CS} = \int_M \left(\kappa(A, dA) + \frac{1}{3}\kappa(A, [A, A]) \right),$$

where A is a gauge field (a connection on a principal G-bundle over M) and κ is a non-degenerate G-invariant symmetric bilinear form on the Lie algebra of G. Clearly, the fermionic field ρ should be regarded as analogous to the bosonic field A, i.e. the Rozansky-Witten model is an odd analogue of Chern-Simons theory. The symplectic form Ω is an analogue of the symmetric bilinear form κ.

This similarity is more than a mere analogy: as shown in [31] the Feynman diagram expansion of the partition function of the Rozansky-Witten model is basically the same as in Chern-Simons theory, with the curvature tensor playing the role of the structure constants of the Lie algebra of G. The partition function of the RW model (i.e. the number assigned to a closed oriented 3-manifold M) was shown in [31] to be a finite-type invariant of M. In the same paper the vector space assigned to a closed oriented 2-manifold of genus g was computed; it turns out to be isomorphic to

$$\bigoplus_p H^p_{\bar\partial}\left(X, \left(\bigwedge T^{1,0}X \right)^{\otimes g} \right)$$

The category associated to a circle turns out to be [29, 30] a 2-periodic version of the derived category of coherent sheaves on X which we denote $D_{\mathbb{Z}_2}(Coh(X))$. Its objects are 2-periodic complexes of coherent sheaves on X, and morphisms are obtained, as usual, by formally inverting quasi-isomorphisms. One way to see this is to consider the Rozansky-Witten model on $S^1 \times \Sigma$, where Σ is a not-necessarily-closed oriented 2-manifold. Kaluza-Klein reduction in this case gives a very simple effective 2d TFT: the so-called B-model with target X which has been studied extensively in connection with Mirror Symmetry [38]. Its category of branes is well understood and is known to be equivalent to the derived category of coherent sheaves [7, 1]. The only difference compared to the usual B-model is that the $\mathbb{Z}$-grading is replaced by $\mathbb{Z}_2$-grading, leading to a 2-periodic version of the derived category.

The fact that the Rozansky-Witten model is $\mathbb{Z}_2$ -graded rather than $\mathbb{Z}$-graded might seem like a technicality, but it becomes crucially important when

one turns to computing the braided monoidal structure on the category assigned to S^1. Both the usual and the 2-periodic derived categories have obvious symmetric monoidal structures given by the derived tensor product. However, the braided monoidal structure on $D_{\mathbb{Z}_2}(Coh(X))$ which arises from the Extended TFT turns out to be not this obvious symmetric monoidal structure but its deformation, which is no longer symmetric [29, 30]. $\mathbb{Z}_2$-grading is important here, because the usual bounded derived category appears not to admit any non-symmetric monoidal deformations (see below).

The braided monoidal deformation of $D_{\mathbb{Z}_2}(Coh(X))$ is a quantum deformation of the obvious monoidal structure, in the same sense that the category of representations of the quantum group is a quantum deformation of the category of representations of the corresponding classical group. That is, corrections to the "obvious" associator arise as quantum corrections in the Feynman diagram expansion of the Rozansky-Witten model. This has been worked out in detail in [30]. The analogue of the Planck constant is the inverse of the symplectic form Ω, i.e. under a rescaling $\Omega \to \lambda\Omega$ the p-th order quantum correction scales like λ^{1-p}.

Finally let us turn to the 2-category that the Rozansky-Witten model assigns to a point, i.e. the 2-category of boundary conditions [17, 18]. Its simplest objects are complex Lagrangian submanifolds of X. There are also more complicated objects which are described by a family of Calabi-Yau manifolds parameterized by points of a complex Lagrangian submanifold Y, or even more abstractly, a family of Calabi-Yau categories (i.e. categories which have the same formal properties as the derived category of coherent sheaves of a Calabi-Yau manifold) over Y. For simplicity we will not discuss these more complicated objects here.

Even for the simplest geometric objects the description of the categories of morphisms turns out quite complicated, so we will discuss only two extreme cases (see [18] for a more general case). First, suppose that two complex Lagrangian submanifolds Y_1 and Y_2 intersect at isolated points, but not necessarily transversely. The category of morphisms in this case is the direct sum of categories corresponding to each intersection point, in the sense that the set of objects is the union of the sets of objects assigned to each point, and spaces of morphisms between objects coming from different intersection points are 0-dimensional vector spaces. Thus it suffices to describe the category corresponding to a single intersection point.

In the neighborhood of the intersection point one may choose complex Darboux coordinates, i.e. choose an identification of the neighborhood with an open subset U of $T^*\mathbb{C}^m$ with its canonical symplectic form. One can always choose this identification so that $Y_1 \cap U$ and $Y_2 \cap U$ are represented by graphs of exact holomorphic 1-forms dW_1 and dW_2 where W_1 and W_2 are holomorphic functions on an open subset V of $\mathbb{C}^m$. The category of morphisms in this case is equivalent to the *category of matrix factorizations* of the function $W_2 - W_1$ [17, 18]. This category was introduced by M. Kontsevich in connection with

Homological Mirror Symmetry and is defined as follows. An object of this category is a $\mathbb{Z}_2$-graded holomorphic vector bundle E on V equipped with a holomorphic endomorphism D of odd degree satisfying $D^2 = W_2 - W_1 + c$, where c is a complex number. Consider two such objects (E_1, D_1, c_1) and (E_2, D_2, c_2). The space of holomorphic bundle maps from E_1 to E_2 is a $\mathbb{Z}_2$-graded vector space equipped with an odd endomorphism D_{12} defined by

$$D_{12}(\phi) = D_2 \cdot \phi - (-1)^{|\phi|}\phi \cdot D_1, \quad \forall \phi \in \mathrm{Hom}(E_1, E_2),$$

where $|\phi| = 0$ or 1 depending on whether ϕ is even or odd. The endomorphism D_{12} satisfies $D_{12}^2 = c_2 - c_1$. The space of morphisms in the category of matrix factorizations from (E_1, D_1, c_1) to (E_2, D_2, c_2) is defined to be the cohomology of the differential D_{12} if $c_1 = c_2$; otherwise it is defined to be zero.

The other relatively simple case is the category of endomorphisms of a complex Lagrangian submanifold Y of X. On the classical level it is equivalent to $D_{\mathbb{Z}_2}(Coh(Y))$. One way to see this is to perform the Kaluza-Klein reduction of the Rozansky-Witten model on an interval $[0, 1]$ with boundary conditions corresponding to Y. On the classical level the effective 2d TFT is a $\mathbb{Z}_2$-graded version of the B-model with target Y, whose category of branes is $D_{\mathbb{Z}_2}(Coh(Y))$. The monoidal structure is the usual derived tensor product. Both the monoidal structure and the category itself are modified by quantum corrections, in general.

3.2. Monoidal deformations of the derived category of coherent sheaves. As far as we know, the theory of deformations of monoidal categories in general, and the derived category of coherent sheaves in particular, has not been systematically developed. A remarkably simple geometric picture of monoidal deformations of the latter category emerges from the study of the Rozansky-Witten model. Consider a complex Lagrangian submanifold Y of a holomorphic symplectic manifold (X, Ω). The functional integral of the Rozansky-Witten model localizes on constant maps, which implies that the category of endomorphisms of the object Y depends only on the formal neighborhood of Y in X. If this formal neighborhood happens to be isomorphic, as a holomorphic symplectic manifold, to the formal neighborhood of the zero section of T^*Y (which we will denote T^*_fY), then one can show that the category of endomorphisms of Y does not receive quantum corrections and therefore is equivalent to $D_{\mathbb{Z}_2}(Coh(Y))$ as a monoidal category.

In general, the formal neighborhood of Y is isomorphic to T^*_fY as a real symplectic manifold, but not as a holomorphic symplectic one. The deviation of the complex structure on T^*_fY from the standard one is described by a $(0, 1)$-form β with values in the graded holomorphic vector bundle

$$\oplus_{p=2}^{\infty}\mathrm{Sym}^p(TY). \tag{2}$$

This $(0, 1)$-form satisfies a Maurer-Cartan-type equation

$$\bar{\partial}\beta + \frac{1}{2}[\beta, \beta] = 0. \tag{3}$$

Here brackets denote wedge product of forms combined with a Lie bracket on sections of $\mathrm{Sym}^\bullet TY$. The Lie bracket comes from the identification of the space of sections of $\mathrm{Sym}^\bullet TY$ with the space of fiberwise-holomorphic functions on $T^*_f Y$ and the Poisson bracket on such functions. Note that because the wedge product of 1-forms is skew-symmetric, the expression $[\beta, \beta']$ is symmetric with respect to the exchange of 1-forms β and β', and therefore $[\beta, \beta]$ need not vanish.

The Rozansky-Witten model provides a map from the space of solutions of the equation (3) to the space of monoidal deformations of $D_{\mathbb{Z}_2}(Coh(Y))$. As usual, there is a group of gauge transformations whose action on the space of solutions is determined by the action of its Lie algebra:

$$a : \beta \mapsto \beta + \bar\partial a + [\beta, a], \tag{4}$$

where a is a section of $\oplus_{p=1}^\infty \mathrm{Sym}^p TY$.

The following natural conjecture was formulated in [17]:

Conjecture. *Let M_Y be the space of solutions of the Maurer-Cartan equation (3) where β is an inhomogeneous form of type $(0, q)$ on Y with odd q with values in the bundle (2). There is a surjective map from M_Y to the space of monoidal deformations of the category $D_{\mathbb{Z}_2}(Coh(Y))$. Two monoidal deformations are equivalent if the corresponding solutions of the Maurer-Cartan equation are related by a gauge transformation whose infinitesimal form is given by (4).*

If true, this conjecture gives an elegant description of all monoidal deformations of $D_{\mathbb{Z}_2}(Coh(Y))$. This description is strikingly similar to the description of all deformations, monoidal or not, of the category $D^b(Coh(Y))$ regarded as an A_∞ category [21]. The latter makes use of a Maurer-Cartan-type equation for a $(0, q)$ form P with values in the graded bundle $\Lambda^\bullet TY$:

$$\bar\partial P + \frac{1}{2}[P, P]_{SN} = 0. \tag{5}$$

Here brackets denote the Schouten-Nijenjuis bracket on polyvector fields, and the total degree of P (that is, the sum of the form degree and the polyvector degree) is even. In particular a holomorphic Poisson bivector, i.e. a section of $\Lambda^2 TY$ satisfying

$$\bar\partial P = 0, \quad [P, P]_{SN} = 0$$

gives rise to a noncommutative deformation of Y. The analog of the holomorphic Poisson bivector in our case is a $(0, q)$-form β with values in $\mathrm{Sym}^2 TY$ satisfying

$$\bar\partial \beta = 0, \quad [\beta, \beta] = 0.$$

The corresponding deformation of the monoidal structure makes the tensor product on $D_{\mathbb{Z}_2}(Coh(Y))$ non-symmetric [17, 18]. Thus one may regard this deformation as a categorification of deformation quantization, and the conjectural relation between the space of solutions of the Maurer-Cartan equation (3)

and the space of monoidal deformations as a categorification of the Formality Theorem of M. Kontsevich [21].

Let us comment on the analogue of the above conjecture in the $\mathbb{Z}$-graded case, i.e. when $D_{\mathbb{Z}_2}(Coh(Y))$ is replaced with $D^b(Coh(Y))$. The latter category can be interpreted as the endomorphism category of a boundary condition in a $\mathbb{Z}$-graded version of the Rozansky-Witten model. This $\mathbb{Z}$-graded version exists if the target manifold X admits a $\mathbb{C}^*$ action with respect to which the holomorphic symplectic form has weight 2. To realize $D^b(Coh(Y))$ as the endomorphism category, we take $X = T^*Y$ with a canonical symplectic form $dpdq$, where p is the fiber coordinate, and define the $\mathbb{C}^*$ action by

$$\lambda : p \mapsto \lambda^2 p, \quad \lambda \in \mathbb{C}^*.$$

That is, the fiber coordinate has weight 2. Accordingly, if we identify the space of sections of $\mathrm{Sym}^p TY$ with the space of functions on T^*Y which are holomorphic degree-p polynomials on the fibers, we should place it in cohomological degree $2p$. A $(0,q)$-form with values in $\mathrm{Sym}^p TY$ will therefore have degree $q + 2p$. From the point of view of the $\mathbb{Z}$-graded Rozansky-Witten model, holomorphic symplectic deformations of T^*Y should be identified with degree-3 solutions of the Maurer-Cartan equation (3). Since $p \geq 2$, the only such solution is the zero one.[5]

The $\mathbb{Z}$-graded Rozansky-Witten model with target T^*Y is particularly simple since one can show that quantum corrections always vanish. This enables one to give a fairly concise description of the 2-category of boundary conditions for this 3d TFT. If we regard a monoidal category $\mathcal{C}$ as a categorification of an associative algebra, then the categorification of a module is a module category over $\mathcal{C}$, .i.e. a category $\mathcal{D}$ on which $\mathcal{C}$ acts by endofunctors. Module categories over a monoidal category form a 2-category. Roughly speaking, the 2-category of boundary conditions for the $\mathbb{Z}$-graded Rozansky-Witten model with target T^*Y is the 2-category of module categories over the monoidal category $D^b(Coh(Y))$. To be more precise, one needs to consider a differential graded version of both the monoidal category $D^b(Coh(Y))$ and module categories over it. The resulting 2-category also appeared in the mathematical papers [4, 33].

Note that the existence of the $\mathbb{Z}$-graded Rozansky-Witten model may be regarded as a "physical reason" for the existence of the derived tensor product on the category $D^b(Coh(Y))$ associated to a complex manifold Y. That is, while the derived tensor product has no physical meaning if one thinks about $D^b(Coh(Y))$ as the category of branes in a 2d TFT (the B-model with target Y), it arises naturally once we realize that the B-model with target Y can be obtained by Kaluza-Klein reduction from a 3d TFT on an interval, with

[5] The restriction to $p \geq 2$ appears because we want the complex structure of Y to be undeformed. If we relax this assumption, then the only allowed β is a $(0,1)$-form with values in TY and satisfying the Maurer-Cartan equation (5). Such β describes a deformation of $D^b(Coh(Y))$ which arises from a deformation of the complex structure on Y. This deformation is obviously monoidal, but not very interesting.

identical boundary conditions at the two endpoints. This 3d TFT is the $\mathbb{Z}$-graded Rozansky-Witten model with target T^*Y, and the boundary condition is the complex Lagrangian submanifold given by the zero section of T^*Y.

4. Topological Gauge Theory in four dimensions and Geometric Langlands Duality

4.1. Electric-magnetic duality and Topological Gauge Theory. Another application of the Extended TFT has to do with the Geometric Langlands Duality. [6] The physical origin of the Geometric Langlands Duality is a conjectural isomorphism between two supersymmetric gauge theories in four dimensions with gauge groups G and ${}^LG$, where ${}^LG$ is the Langlands-dual of G.[7] This isomorphism is known as Montonen-Olive duality [27] and holds for gauge theories with maximal supersymmetry [39, 28]. There are many computations verifying particular implications of the Montonen-Olive conjecture, but no general proof. In the case $G = U(1) = {}^LG$, the Montonen-Olive duality reduces to electric-magnetic duality, i.e. the transformation which exchanges electric and magnetic fields. Electric-magnetic duality is well known to be a symmetry of the $U(1)$ gauge theory both on the classical and quantum levels. The Montonen-Olive conjecture can be regarded as a far-reaching nonabelian generalization of this fact.

To connect the Montonen-Olive conjecture to Geometric Langlands Duality the first step is to replace supersymmetric gauge theories with much simpler topological field theories [20]. This is achieved by means of a procedure called *twisting* [35]. Roughly speaking, one redefines the stress-energy tensor of the theory, so that it becomes Q-exact with respect to a certain nilpotent odd vector field Q on the supermanifold of all field configurations. For historical reasons, Q is known as the BRST operator. Simultaneously one restricts observables (i.e. functions on the supermanifold of field configurations) to those which are annihilated by Q. After twisting both the supersymmetric gauge theory with gauge group G and its Montonen-Olive dual with gauge group ${}^LG$ one gets a pair of isomorphic 4d TFTs.

In fact, the situation is more complicated than that. First of all, the maximally supersymmetric gauge theory in four dimensions admits three inequivalent twists differing both by the choice of Q and by the required modification of the stress-energy tensor [34]. Following [20] we will focus on the twist which was first considered by N. Marcus [26] and is nowadays called the GL-twist. Second, the GL-twisted gauge theory has two supercommuting BRST operators Q_l and

[6] The relationship between ETFT and Geometric Langlands Duality is also discussed in [5].

[7] We recall that the Langlands dual of compact simple Lie group G is a compact simple Lie group ${}^LG$ whose maximal torus is isomorphic to the dual of the maximal torus of G.

Q_r, and one can take any linear combination of this as the BRST operator which must annihilate the observables:

$$Q = uQ_l + vQ_r, \quad u, v \in \mathbb{C}, \quad |u|^2 + |v|^2 > 0.$$

The overall normalization of Q does not affect the theory, so the GL-twisted theory is really a family of 4d TFTs parameterized by points of $\mathbb{CP}^1$ [20]. We will identify $\mathbb{CP}^1$ with a one-point compactification of the complex plane and will label a particular TFT by a parameter $t \in \mathbb{C} \cup \{\infty\}$.

It turns out that Montonen-Olive duality not only replaces G with ${}^LG$ but also acts on the parameter t [20]. Geometric Langlands duality arises from a particular instance of the Montonen-Olive duality which maps $t = i$ to $t = 1$.[8]

4.2. From Topological Gauge Theory to Geometric Langlands Duality. We can now attempt to extract some mathematical consequences of the Montonen-Olive conjecture. Let us fix a compact simple Lie group G and let C be a closed oriented 2-manifold. The 4d Topological Gauge Theory with gauge group G assigns to C a family of $\mathbb{C}$-linear categories parameterized by t; we will denote a member of this family by $\mathsf{F}(G, t, C)$. The Montonen-Olive conjecture implies that there is an equivalence of categories

$$\mathsf{F}(G, i, C) \simeq \mathsf{F}({}^LG, 1, C).$$

It remains to understand the categories involved. This turns out to be rather nontrivial. The category $\mathsf{F}(G, t, C)$ is the category of branes for the 2d TFT obtained by Kaluza-Klein reduction of the 4d Topological Gauge Theory on C. For $g > 1$ this 2d TFT was analyzed in [20]. It was shown that for $t = i$ the 2d TFT is a B-model whose target is the moduli space $\mathcal{M}_{flat}(G_{\mathbb{C}}, C)$ of flat $G_{\mathbb{C}}$-connections on C. Here $G_{\mathbb{C}}$ is the complexification of G. Accordingly, for $t = i$ the category of branes of the 2d TFT is the derived category of coherent sheaves on $\mathcal{M}_{flat}(G_{\mathbb{C}}, C)$. For $t = 1$ the 2d TFT is a different topological sigma-model (the A-model) whose target space is a symplectic manifold $\mathcal{M}^{symp}_{flat}(G_{\mathbb{C}}, C)$. This manifold is diffeomorphic to $\mathcal{M}_{flat}(G_{\mathbb{C}}, C)$, with an exact symplectic form given by

$$\omega = \int_C \kappa(\delta A \wedge \delta\phi),$$

where A is the real part of the flat $G_{\mathbb{C}}$-connection on C, ϕ is its imaginary part, and κ is the Killing metric on the Lie algebra of G. The category of branes for an A-model is the so-called Fukaya-Floer category [13] whose objects are Lagrangian submanifolds of the target space and whose morphisms are defined by means of the Lagrangian Floer homology. Thus Montonen-Olive duality implies that the derived category of coherent sheaves on $\mathcal{M}_{flat}(G_{\mathbb{C}}, C)$ is equivalent to the Fukaya-Floer category of $\mathcal{M}^{symp}_{flat}({}^LG_{\mathbb{C}}, C)$.

[8] More generally, one gets what is known as Quantum Geometric Langlands.

The usual statement of the Geometric Langlands Duality is somewhat different. Instead of the Fukaya-Floer category of $\mathcal{M}^{symp}_{flat}({}^L G_{\mathbb{C}}, C)$ it involves the derived category of D-modules over the moduli stack of holomorphic ${}^L G_{\mathbb{C}}$-bundles over C. It was shown in [20] that there is a functor from the former to the latter, but it is not clear at the time of writing why this functor should be an equivalence.

If C has genus zero, the effective 2d TFT one obtains by Kaluza-Klein reduction is rather different. For $t = i$ it was shown in [19] that the 2d TFT is a $G_{\mathbb{C}}$-equivariant B-model whose target is the Lie algebra $\mathfrak{g}_{\mathbb{C}}$ of $G_{\mathbb{C}}$ placed in cohomological degree 2. That is, the target is a purely even graded manifold which we denote $\mathfrak{g}_{\mathbb{C}}[2]$. The group $G_{\mathbb{C}}$ acts on $\mathfrak{g}_{\mathbb{C}}$ by the adjoint representation. The category of branes for this 2d TFT is the $G_{\mathbb{C}}$-equivariant derived category of coherent sheaves on $\mathfrak{g}_{\mathbb{C}}[2]$ [19]. For $t = 0$ and genus zero the 2d TFT has not been analyzed thoroughly yet. From the mathematical viewpoint, one expects that Geometric Langlands Duality relates the category $D^b_{G_{\mathbb{C}}}(Coh(\mathfrak{g}[2]))$ and the ${}^L G$-equivariant constructible derived category of sheaves on the loop group of ${}^L G$ [6]. It would be interesting to see if the latter category emerges as the category of branes in a 2d TFT.

Instead of studying categories which the 4d TFT assigns to a closed oriented 2-manifold, we may consider 2-categories assigned to a circle. For a fixed gauge group G we get a family of these 2-categories parameterized by t. Let us denote a member of this family by $\mathsf{F}(G, t, S^1)$. The Montonen-Olive conjecture implies an equivalence of 2-categories:

$$\mathsf{F}(G, i, S^1) \simeq \mathsf{F}({}^L G, 1, S^1). \tag{6}$$

Moreover, as mentioned above, both 2-categories are supposed to have braided monoidal structure, and the equivalence is supposed to be compatible with them.

To understand the 2-categories $\mathsf{F}(G, t, S^1)$ one needs to study the Kaluza-Klein reduction of the 4d Topological Gauge Theory on a circle. The corresponding 3d TFT has been analyzed in [19]. It turns out that at $t = i$ the 3d TFT is a $G_{\mathbb{C}}$-equivariant version of the $\mathbb{Z}$-graded Rozansky-Witten model with target $T^*G_{\mathbb{C}}$, where $G_{\mathbb{C}}$ acts on $T^*G_{\mathbb{C}}$ by conjugation. The corresponding 2-category of branes is, roughly speaking, the 2-category of module categories over the $G_{\mathbb{C}}$-equivariant derived category of coherent sheaves over $G_{\mathbb{C}}$, regarded as a monoidal category. A typical object of this 2-category is a family of categories over $G_{\mathbb{C}}$ with an action of $G_{\mathbb{C}}$ which lifts the action of $G_{\mathbb{C}}$ on itself by conjugation. For $t = 1$ the 3d TFT has not been studied thoroughly, but it appears that its objects are module categories over a G-equivariant version of the Fukaya-Floer category of T^*G, regarded as a monoidal category.[9]

[9] The monoidal structure on this category is less obvious than for the analogous category at $t = i$.

5. Open Questions

It is plausible that the correct mathematical formalism for Extended Topological Field Theories is provided by the theory of (∞, n)-categories developed by J. Lurie [24, 23]. More precisely, what seems most relevant is the linear case of this theory where the set of n-morphisms has the structure of a differential graded vector space.[10] Specifically, one expects that to every n-dimensional TFT one can attach a linear $(\infty, n-1)$-category whose objects are boundary conditions, whose 1-morphisms are boundary-changing operators supported on submanifolds of the boundary of codimension 1, etc. Compositions of k-morphisms arise from the fusion of the physical operators supported on submanifolds of the boundary of codimension k. It would be interesting to understand whether the definition of the (∞, n)-category captures the properties of fusion expected on physical grounds.

Extended Topological Field Theory provides a new viewpoint on the Geometric Langlands Program. The most powerful statement implied by the topologically twisted version of the Montonen-Olive conjecture is the equivalence of 3-categories which the 4d Topological Gauge Theories with gauge groups G and ${}^L G$ assign to a point. From the physical viewpoint, these are 3-categories of boundary conditions. Some examples of boundary conditions for maximally supersymmetric gauge theories and their Montonen-Olive duals have been constructed in [14, 15], but we are still far from understanding the nature of this 3-category.

One categorical level down, it would be very interesting to study the rich structure on the 2-category $\mathsf{F}(G, t, S^1)$ which the 4d Topological Gauge Theory assigns to S^1. We already mentioned that it has a braided monoidal structure and an identity object. The monoidal structure arises from the bordism from $S^1 \sqcup S^1$ to S^1 depicted in fig. 8. Another way to draw it is shown in fig. 9; we will call it the "pants" bordism. Similarly, the identity object arises from a disc regarded as a bordism between the empty 1-manifold and S^1 (the "cup" bordism, see fig. 9). Further, the 2-category $\mathsf{F}(G, t, S^1)$ is expected to be *rigid*, i.e. there should be a 2-functor e from $\mathsf{F}(G, t, S^1) \times \mathsf{F}(G, t, S^1)$ to the 2-category of linear categories, and a 2-functor ι in the opposite direction satisfying some compatibility conditions. These 2-functors both arise from a cylinder regarded either as a bordism from $S^1 \sqcup S^1$ to the empty 1-manifold (the "downward plumbing fixture", see fig. 9), or as a bordism from the empty 1-manifold to $S^1 \sqcup S^1$ (the "upward plumbing fixture", see fig. 9). All these 2-functors should satisfy various compatibility conditions arising from the fact that although one can glue a given oriented 2-manifold with boundaries from these four building blocks ("pants", "cup" and two "plumbing fixtures") in many different ways, the equivalence class of 2-functors corresponding to this 2-manifold must be

[10] From the physical viewpoint, linearity arises from the fact that the quantum-mechanical space of states is a (graded) vector space.

well-defined. One can summarize the situation by saying that $\mathsf{F}(G, t, S^1)$ is a rigid braided monoidal 2-category. A related viewpoint on the Geometric Langlands Duality (not using the language of Extended Topological Field Theory) is proposed in [12].

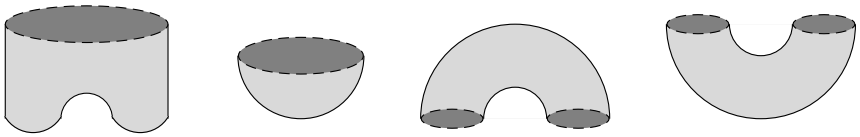


Figure 9. Basic 2d bordisms: pants, cup, downward plumbing fixture, upward plumbing fixture.

Montonen-Olive duality predicts that $\mathsf{F}(G, i, S^1)$ and $\mathsf{F}({}^L G, 1, S^1)$ are equivalent as rigid braided monoidal 2-categories. This statement should imply the statement of the usual Geometric Langlands Duality in the following way. Given any closed oriented 2-manifold C we may represent it as a result of gluing the four building blocks shown in fig. 9. Extended TFT in four dimensions assigns to every building block a 2-functor as described above, and therefore assigns to the whole C a 2-functor from the 2-category of linear categories to itself. Axioms of Extended TFT ensure that the equivalence class of this 2-functor is independent of the way one cut C into pieces. The category $\mathsf{F}(G, t, C)$ can be thought of as the result of applying this 2-functor to the category of vector spaces. The rigid braided monoidal equivalence of $\mathsf{F}(G, i, S^1)$ and $\mathsf{F}({}^L G, 1, S^1)$ then implies the equivalence of categories $\mathsf{F}(G, i, C)$ and $\mathsf{F}({}^L G, 1, C)$.

Acknowledgments

I would like to thank David Ben-Zvi, Dmitri Orlov, and Lev Rozansky for comments on the draft of the talk.

References

[1] P. Aspinwall et al., "Dirichlet branes and Mirror Symmetry," *Clay Mathematics Monographs*, **4**. American Mathematical Society, 2009.

[2] M. Atiyah, "Topological Quantum Field Theories," Inst. Hautes Etudes Sci. Publ. Math. **68**, 175 (1989).

[3] J. C. Baez and J. Dolan, "Higher dimensional algebra and topological quantum field theory," J. Math. Phys. **36**, 6073 (1995) [arXiv:q-alg/9503002].

[4] D. Ben-Zvi, J. Francis and D. Nadler, "Integral Transforms and Drinfeld Centers in Derived Algebraic Geometry," arXiv:0805.0157 [math.AG].

[5] D. Ben-Zvi and D. Nadler, "The Character Theory of a Complex Group," arXiv:0904.1247 [math.RT].

[6] R. Bezrukavnikov and M. Finkelberg, "Equivariant Satake category and Kostant-Whittaker reduction," Mosc. Math. J. **8**, 39 (2008) [arXiv:0707.3799 [math.RT]]

[7] M. R. Douglas, "D-branes, categories and N = 1 supersymmetry," J. Math. Phys. **42**, 2818 (2001) [arXiv:hep-th/0011017].

[8] D. S. Freed, "Higher algebraic structures and quantization," Commun. Math. Phys. **159**, 343 (1994) [arXiv:hep-th/9212115].

[9] D. S. Freed, "Quantum groups from path integrals," arXiv:q-alg/9501025.

[10] D. S. Freed, " Remarks on Chern-Simons theory," Bull. Amer. Math. Soc. **46**, 221 (2009).

[11] D. S. Freed, M. J. Hopkins, J. Lurie and C. Teleman, "Topological Quantum Field Theories from Compact Lie Groups," arXiv:0905.0731 [math.AT].

[12] E. Frenkel and D. Gaitsgory, "Local Geometric Langlands Correspondence and Affine Kac-Moody algebras," arXiv:math/0508382 [math.RT].

[13] K. Fukaya, Y-G. Oh, H. Ohta, K. Ono, "Lagrangian intersection Floer theory: anomaly and obstruction. Part I." American Mathematical Society, 2009.

[14] D. Gaiotto and E. Witten, "Supersymmetric Boundary Conditions in N=4 Super Yang-Mills Theory," arXiv:0804.2902 [hep-th].

[15] D. Gaiotto and E. Witten, "S-Duality of Boundary Conditions In N=4 Super Yang-Mills Theory," Adv. Theor. Math. Phys. **13**, 721 (2010) [arXiv:0807.3720 [hep-th]].

[16] M. Kapranov and V. Voevodsky, "2-categories and Zamolodchikov tetrahedra equations," in: *Algebraic groups and their generalizations: quantum and infinite-dimensional methods*, 177–259, Proc. Sympos. Pure Math, **56**, Part 2. American Mathematical Society, 1994.

[17] A. Kapustin, L. Rozansky and N. Saulina, "Three-dimensional topological field theory and symplectic algebraic geometry I," Nucl. Phys. B **816**, 295 (2009) [arXiv:0810.5415 [hep-th]].

[18] A. Kapustin and L. Rozansky, "Three-dimensional topological field theory and symplectic algebraic geometry II," arXiv:0909.3643 [math.AG].

[19] A. Kapustin, K. Setter and K. Vyas, "Surface operators in four-dimensional topological gauge theory and Langlands duality," arXiv:1002.0385 [hep-th].

[20] A. Kapustin and E. Witten, "Electric-magnetic duality and the Geometric Langlands Program," Commun. Number Theory Phys. **1**, 1 (2007).

[21] M. Kontsevich, "Deformation quantization of Poisson manifolds, I," Lett. Math. Phys. **66**, 157 (2003) [arXiv:q-alg/9709040].

[22] R. Lawrence, "An introduction to topological field theory," in: *The Interface of Knots and Physics (San Francisco, 1995)*, Proc. Sympos. Appl. Math. **51**, 89. American Mathematical Society, 1996.

[23] J. Lurie, "$(\infty, 2)$-Categories and the Goodwillie Calculus I," arXiv:0905.0462 [math.CT]

[24] J. Lurie, "On the classification of Topological Field Theories," in: *Current developments in mathematics, 2008*, 129–280. Int. Press, Sommerville, MA, 2009 [arXiv:0905.0465 [math.AT]]

[25] S. MacLane, "Categories for the working mathematician." Springer, 1971.

[26] N. Marcus, "The other topological twisting of N=4 Yang-Mills," Nucl. Phys. B **452**, 331 (1995) [arXiv:hep-th/9506002].

[27] C. Montonen and D. I. Olive, "Magnetic Monopoles As Gauge Particles?," Phys. Lett. B **72**, 117 (1977).

[28] H. Osborn, "Topological Charges For N=4 Supersymmetric Gauge Theories And Monopoles Of Spin 1," Phys. Lett. B **83**, 321 (1979).

[29] J. Roberts, "Rozansky-Witten theory," arXiv:math/0112209 [math.QA]

[30] J. Roberts and S. Willerton, "On the Rozansky-Witten weight system," arXiv:math/0602653 [math.DG]

[31] L. Rozansky and E. Witten, "Hyper-Kaehler geometry and invariants of three-manifolds," Selecta Math. **3**, 401 (1997) [arXiv:hep-th/9612216].

[32] A. S. Schwarz, "The Partition Function Of A Degenerate Functional," Commun. Math. Phys. **67**, 1 (1979).

[33] B. Toen, G. Vezzosi, "A note on Chern character, loop spaces, and derived algebraic geometry," arXiv:0804.1274 [math.AG].

[34] C. Vafa and E. Witten, "A Strong coupling test of S duality," Nucl. Phys. B **431**, 3 (1994) [arXiv:hep-th/9408074].

[35] E. Witten, "Topological Quantum Field Theory," Commun. Math. Phys. **117**, 353 (1988).

[36] E. Witten, "Topological Sigma Models," Commun. Math. Phys. **118**, 411 (1988).

[37] E. Witten, "Quantum field theory and the Jones polynomial," Commun. Math. Phys. **121**, 351 (1989).

[38] E. Witten, "Mirror manifolds and topological field theory," arXiv:hep-th/9112056.

[39] E. Witten and D. I. Olive, "Supersymmetry Algebras That Include Topological Charges," Phys. Lett. B **78**, 97 (1978).

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Origins of Diffusion

Antti Kupiainen*

Abstract

We consider a dynamical system consisting of subsystems indexed by a lattice. Each subsystem has one conserved degree of freedom ("energy") the rest being uniformly hyperbolic. The subsystems are weakly coupled together so that the sum of the subsystem energies remains conserved. We prove that the long time dynamics of the subsystem energies is diffusive.

Mathematics Subject Classification (2010). Primary 37L60; Secondary 82C05.

Keywords. Coupled map lattices, diffusion, hydrodynamic limit, renormalization group

1. Diffusion from Conservative Dynamics

One of the fundamental problems in deterministic dynamics is to understand the microscopic origin of dissipation and diffusion. On a microscopic level a physical system such as a fluid or a crystal can be described by a Schrödinger or a Hamiltonian dynamical system with a macroscopic number of degrees of freedom. Although the microscopic dynamics is reversible in time one expects dissipation to emerge in large spatial and temporal scales e.g. in the form of diffusion of heat or concentration of particles.

To fix ideas, consider a Hamiltonian dynamical system i.e. a Hamiltonian flow on a symplectic manifold M. For the present purpose it suffices to consider $M = \mathbb{R}^{2n}$ with position and momentum coordianates $q, p \in \mathbb{R}^n$. The Hamiltonian flow $\phi_t \in \mathrm{Diff}\, M$ generated by the vector field $(\partial_p H, -\partial_q H)$ where $H : M \to \mathbb{R}$ is the Hamiltonian or energy function preserves the energy

$$H \circ \phi_t = H$$

i.e. the flow preserves the constant energy sets $M_E = \{(q,p) : H(q,p) = E\}$.

*Supported by European Research Council and Academy of Finland
Helsinki University, Department of Mathematics, P.O.Box 68, 00014, Helsinki, Finland. E-mail: ajkupiai@cc.helsinki.fi.

On the other hand, the simplest diffusion process is given by the heat equation

$$\partial_t E(t,x) = \kappa \Delta E(t,x) \tag{1}$$

and the associated semigroup $\psi_t = e^{\kappa t \Delta}$. Unlike for the reversible ϕ_t where $\phi_{-t} = \phi_t^{-1}$, ψ_t has no inverse and describes dissipation. Physically, the energy function $E(t,x)$ describes a macroscopic energy density i.e. a coarse grained function of microscopic dynamical variables, the positions and momenta of the underlying Hamiltonian dynamics. The question we wish to pose is how does this dissipative dynamics ψ_t arise from the conservative one ϕ_t.

A concrete physical system where diffusion occurs is a fluid. In classical mechanics this is microscopically modeled by a Hamiltonian system whose flow gives the trajectories of the fluid particles $(q_i(t), p_i(t)) \in \mathbb{R}^3 \times \mathbb{R}^3$, $i = 1 \dots N$. A typical Hamiltonian function is given by

$$H(q,p) = \sum_i \frac{p_i^2}{2m} + \sum_{ij} V(q_i - q_j) \tag{2}$$

consisting of the kinetic energy of the particles of mass m and a pair potential energy of interaction of the particles. Let the energy of the i:th particle be defined as

$$e_i = \frac{p_i^2}{2m} + \tfrac{1}{2} \sum_{j \neq i} V(q_i - q_j) \tag{3}$$

so that $H = \sum_i e_i$. We can then define the energy density as the distribution

$$E(t,x) = \sum_i e_i \delta_{q_i(t)}(x) \tag{4}$$

where δ_q is the Dirac mass at q. Since $\int E(t,x)dx = \sum_i e_i = H$ and $\dot{H} = 0$ one concludes

$$\dot{E}(t,x) = \nabla \cdot J(t,x) \tag{5}$$

for a certain distribution, the energy current, depending on $q(t), p(t)$. Eq. (5) is a local conservation law deduced from the global energy conservation. In the case of the fluid, there are two other similar local conservation laws related to global momentum and particle number conservation laws. This leads to a richer set of macroscopic laws in the case of the fluid than the diffusion equation for the energy (in particular these include the Navier-Stokes equations).

2. Coupled Oscillators

Thus, to understand the origins of diffusion one should look for systems with just one local conservation law eq. (5). There has been a lot of work in recent

years around these questions in the context of **coupled dynamics** i.e. dynamical systems consisting of elementary systems indexed by a d-dimensional lattice $\mathbb{Z}^d$. The total energy E of the system is a sum $\sum_x E_x$ of energies E_x which involve the dynamical variables of the system at lattice site x and nearby sites. The physical situation to keep in mind is then thermal conduction in a crystal lattice (i.e. a solid).

Two types of systems have been considered. In the first case at each lattice site we have an oscillator and the oscillators at neighboring sites are coupled together. Typically one considers the system where the forces are weakly anharmonic. In the second case at each lattice site one puts a chaotic system and weakly couples the neighboring systems. Let us start with the former case.

The setup resembles that of the fluid above, but now the "particle" positions q_x are indexed by the lattice, $x \in \Lambda \subset \mathbb{Z}^d$ where Λ is a finite subset, say a cube, and they describe the deviation of an atom from its equilibrium position at x. A simple classical mechanical model for this is a system of coupled oscillators

$$H_\Lambda(q,p) = \sum_{x\in\Lambda}\left(\frac{p_x^2}{2m} + U(q_x)\right) + \sum_{|x-y|=1} V(q_x - q_y) \tag{6}$$

where U is a *pinning* potential which we assume tending to infinity as $|q| \to \infty$. The potential V describes interaction of the atoms in nearest neighbor lattice sites and is taken attractive. A challenging model is obtained already by taking

$$V(q) = q^2, \quad U(q) = q^2 + \lambda|q|^4 \tag{7}$$

and further simplifying by taking $q_x \in \mathbb{R}$ instead of $\mathbb{R}^d$. Then a lattice version of eq. (5) holds with the current given by

$$J_\mu(x) = -\tfrac{1}{2}(p_{x+\mu} + p_x)V'(q_{x+\mu} - q_x). \tag{8}$$

In what sense should we expect the conservative dynamics (5) give rise to a diffusive one as in eq. (1)? The answer is that this should happen for *typical* initial conditions $(q(0), p(0)) \in M_\Lambda$ with respect to a specific measure on the phase space $M_\Lambda := \mathbb{R}^{2|\Lambda|}$ and under a proper *scaling limit*.

Recall first that the Hamiltonian dynamics preserves the Lebesgue measure m_Λ on M_Λ. Since also H_Λ is preserved so is the Gibbs measure (or *equilibrium measure*)

$$\mu_{\beta\Lambda} = Z_\Lambda^{-1} e^{-\beta H_\Lambda} m_\Lambda$$

where $\beta > 0$ as well as its (thermodynamic) limit $\mu_\beta = \lim_{\Lambda\to\mathbb{Z}^d} \mu_{\beta\Lambda}$. Let us now replace the (inverse) temperature parameter β by a spatially varying one. Let $b \in C_0^\infty(\mathbb{R}^d)$ and $\beta > \|b\|_\infty$. Write as in the fluid case

$$H_\Lambda = \sum_{x\in\Lambda} e_x$$

e_x describing the energy contributed by the oscillator at x. Pick a scaling parameter $L \in \mathbb{N}$ and set $\beta_L(x) = \beta + b(x/L)$. Let $\mu^{(L)}$ be the thermodynamic limit of the measure

$$Z_{L,\Lambda}^{-1} e^{-\sum_{x\in\Lambda} \beta_L(x) e_x} m_\Lambda.$$

Construction of this limit poses no problems if $\lambda \geq 0$ in eq. (7) is small enough. $\mu^{(L)}$ is *not* invariant under the dynamics which maps it to $\mu_t^{(L)} = \mu^{(L)} \circ \phi_t^{-1}$. However, one expects that as $t \to \infty$ there is *return to equilibrium* i.e. $\mu_t^{(L)} \to \mu_\beta$. The diffusion equation is expected to govern this process in the following sense.

Let $f \in C_0^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$ and consider the random variables

$$e_L(f) = L^{-d-2} \sum_{(t,x)\in\mathbb{Z}_+\times\mathbb{Z}^d} f(t/L^{-2}, x/L) e_x(q(t), p(t)). \tag{9}$$

The statement of the *hydrodynamic limit* is then: with probability one in the sequence of measures $\mu^{(L)}$, $e_L(f)$ converges to $\int f(t,x)E(t,x)dtdx$ where E is the solution to the nonlinear diffusion equation

$$\partial_t E = \nabla \cdot (\kappa(E)\nabla E) \tag{10}$$

where $\kappa(E)$ is a smooth positive function. The initial condition $E(0,\cdot)$ is determined by the function b. Thus upon coarse graining and scaling the equation (5) turns to eq. (10), *almost surely* in the initial conditions of the underlying microscopic variables.

The proof of the hydrodynamic limit in our model is beyond present mathematical techniques. The existing techniques require the presence of plenty of noise in the system. A simpler problem would be to establish the *kinetic limit*. This is a *weak anharmonicity* limit. We replace λ in eq. (7) by $\lambda/\sqrt{L}$ and and consider the measures $\mu_{Lt}^{(L)}$. As $L \to \infty$ we expect these measures to become gaussian whose covariance upon spatial scaling satisfies a Boltzman equation. More precisely, denote (q_x, p_x) by $\phi(x)$. Then it is conjectured that

$$\lim_{L\to\infty} \int \phi(Lx+y)\phi(Lx-y)\mu_{Lt}^{(L)}(d\phi) = G(t,x,y) \tag{11}$$

exists and the Fourier transform of $G(t,x,y)$ in y, $\hat{G}(t,x,k)$ satisfies the so called phonon Boltzman equation

$$\partial_t \hat{G}(t,x,k) + \nabla\omega(k) \cdot \nabla \hat{G}(t,x,k) = I(\hat{G}(t,x,\cdot)) \tag{12}$$

where I is a nonlinear integral operator and $\omega(k)^2$ is the Fourier transform of the lattice operator $2(-\Delta + 1)$, see [1]. Proof of these statements is still open and a considerable challenge (for some progress see [2]). Derivation of a hydrodynamic equation of the type (10) from the Boltzman equation (12) has been carried out [3], see also [4] where an attempt to go beyond the kinetic limit was carried out.

3. Coupled Chaotic Flows

A second class of models deals with a complementary situation of *weakly coupled chaotic* systems [5], [6], [7]. The setup is as follows. Let (M, H) be a Hamiltonian system i.e. M is a symplectic manifold and $H : M \to \mathbb{R}$. Let, for each $x \in \mathbb{Z}^d$ (M_x, H_x) be a copy of (M, H). Let $h : M \times M \to \mathbb{R}$ and for each $x, y \in \mathbb{Z}^d$, $|x - y| = 1$ let $h_{xy} : M_x \times M_y \to \mathbb{R}$ be a copy of h. Let $\Lambda \subset \mathbb{Z}^d$ be finite and $M_\Lambda = \times_{x\in\Lambda} M_x$. The coupled flow is the one on M_Λ generated by the Hamiltonian

$$H_\Lambda = \sum_{x\in\Lambda} H_x + \sum_{|x-y|=1} \lambda h_{xy}. \tag{13}$$

Of course, the coupled oscillators of the previous section are of this form. There, the system (M, H) is integrable, and the diffusive dynamics is the consequence of coupling and anharmonicity. In the present discussion we wish to take (M, H) chaotic. Examples are Anosov systems or billiard systems. E.g. in the former case the flow on M generated by H has $\dim M - 2$ non-zero Lyapunov exponents and two vanishing ones corresponding to the Hamiltonian vector field and ∇H.

When the coupling parameter λ is zero (M_Λ, H_Λ) has $2|\Lambda|$ vanishing Lyapunov exponents. For $\lambda \neq 0$ one expects that for a large class of perturbations h the only constant of motion is H_Λ and the system has only two vanishing exponents. However, zero should be near degenerate for the Lyapunov spectrum and these long time scale motions should be at the origin to diffusion in the $\Lambda \to \mathbb{Z}^d$ limit.

Rigorous results on such Hamiltonian systems are rare: in [5] ergodicity is proved in a one dimensional model. However, it seems very difficult to get hold of the Lyapunov spectrum and it is far from obvious how such knowledge would turn into a proof of diffusion in these systems. I want to argue that a more fruitful approach is to study the local energy conservation law (5) and try to show that the chaotic degerees of freedom act there like a noise that redistributes locally the energy. To probe such an idea it is useful to turn to a discrete time version of our model i.e. to study iteration of a map rather than a flow.

4. Coupled Chaotic Maps

A discrete time version of the coupled flow setup of the previous section is called a Coupled Map Lattice (CML). Now the local dynamical system is a pair (M, f) where M is a manifold and $f : M \to M$. Again for each $x \in \mathbb{Z}^d$ (M_x, f_x) is a copy of (M, f) and (M_Λ, f_Λ) with $f_\Lambda = \times_{x\in\Lambda} f_x$ is the product dynamics. The CML dynamics is a suitable local perturbation of the product dynamics.

Our choice of M and f is motivated by the coupled chaotic flows discussed before. A discrete time version (say given by a Poincare map) of a billiard

or Anosov flow has one vanishing Lyapunov exponent corresponding to the conserved energy and the remaining ones nonzero. We model such a situation by taking for the local dynamics the manifold of form $M = \mathbb{R}_+ \times N$ with N another manifold. Let us denote the variables at the lattice site $x \in \mathbb{Z}^d$ by $(E(x), \theta(x)) \in \mathbb{R}_+ \times N$. We call the non-negative variables E energy and postulate them to be conserved under the local dynamics:

$$(E(x), \theta(x)) \to (E(x), g(\theta(x), E(x))) \tag{14}$$

for each $x \in \mathbb{Z}^d$.

$\theta \in N$ are the fast, chaotic variables. In the billiard case the dynamical system $\theta \to g(\theta, E)$ is uniformly hyperbolic for any fixed E. We will model this situation by taking $g(\theta, E) = g(\theta)$ a fixed chaotic map, independent of E. Examples are $N = \mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ and g an expansive circle map, e.g $g(\theta) = 2\theta$ and $N = \mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ and g a hyperbolic toral automorphism.

We should stress that the E independence is the most serious simplification in this setup. In a realistic Hamiltonian system, such as the billiards the E dependence of g can not be ignored. Indeed, it is obvious that as $E \to 0$ the Lyapunov exponents of $g(\cdot, E)$ also tend to zero since E sets the time scale.

The CML dynamics is a perturbation of the local dynamics (14). Let us use the same notation $(E, \theta) \in M_\Lambda = \mathbb{R}_+^\Lambda \times N^\Lambda$. Then $F : M_\Lambda \to M_\Lambda$ is written as

$$F(x, E, \theta) = (E(x) + f(x, E, \theta), g(\theta(x)) + h(x, \theta)). \tag{15}$$

Here f and h are small local functions of (E, θ) i.e. they depend weakly on $(E(y), \theta(y))$ for $|x - y|$ large as we will specify later.

f is however constrained by the requirement that the *total energy* $\sum_x E(x)$ is conserved. This follows if

$$\sum_x f(x, E, \theta) = 0$$

for all E, θ. A natural way to guarantee this is to consider a "vector field" $\mathbf{J}(x) = \{J^\mu(x)\}_{\mu=1,\ldots,d}$ and take

$$f(x, E, \theta) = (\nabla \cdot \mathbf{J})(x, E, \theta) := \sum_\mu (J^\mu(x + e_\mu, E, \theta) - J^\mu(x, E, \theta)) \tag{16}$$

With these definitions we arrive at the time evolution

$$E(t+1, x) = E(t, x) + \nabla \cdot \mathbf{J}(x, E(t), \theta(t)) \tag{17}$$

$$\theta(t+1, x) = g(\theta(t, x)) + h(x, \theta(t))). \tag{18}$$

Note that (17) is a natural discrete space time version of (5). Let us discuss this iteration from a general perspective before making more specific assumptions of the perturbations.

5. Fast Dynamics

The iteration (18) of the chaotic variables is autonomous. We shall assume the perturbation h is C^1 with the following locality property

$$|\partial_{\theta(y)} h(x,\theta)| \leq \epsilon e^{-a|x-y|} \tag{19}$$

and Hölder continuity property

$$|\partial_{\theta(y)} h(x,\theta) - \partial_{\theta(y)} h(x,\theta')| \leq \epsilon \sum_z e^{-a(|x-y|+|x-z|)} |\theta(z) - \theta'(z)|. \tag{20}$$

These properties guarantee [8] that the θ-dynamics is *space-time mixing*. This means that the dynamics is defined in the $\Lambda \to \mathbb{Z}^d$ limit and it has a unique Sinai-Ruelle-Bowen measure μ on the cylinder sets of $N^{\mathbb{Z}^d}$ which satisfies

$$\mathbb{E}(F(\theta(t,x))G(\theta(0,y))) - \mathbb{E}(F(\theta(t,x))\mathbb{E}G(\theta(0,y)) \leq Ce^{-c(t+|x-y|)} \tag{21}$$

for Hölder continuous functions F and G. Here $\mathbb{E}$ denotes expectation in μ.

We conclude that sampling $\theta(0,\cdot)$ with μ makes $\theta(t,x)$ *random variables* which are exponentially weakly correlated at distinct space time points. Therefore $\theta_x(t)$ acts as a *random environment* for the slow variable dynamics (17).

6. Quenched Diffusion

The previous discussion shows that we can view the current $\mathbf{J}(x,E,\theta(t))$ in the slow variable dynamics (17) as a random field $\mathbf{J}(t,x,E)$ which is exponentially weakly correlated in space and time. We may thus rephrase the problem of deriving diffusion in deterministic dynamics as that of *quenched diffusion in random dynamics*. We want to show that the random dynamical system

$$E(t+1,x) \quad = \quad E(t,x) + \nabla \cdot \mathbf{J}(t,x,E(t)) := \Phi(t,x,E(t)) \tag{22}$$

has a diffusive hydrodynamical limit *almost surely* with respect to the SRB measure μ. Let us inquire how this should come about and then list the assumptions we need for the actual proof.

Consider first the *annealed* problem, i.e. averaged equation (22):

$$E_x(t+1) - E_x(t) = \nabla \cdot \mathbb{E}[J(t,x,E(t))] := \nabla \cdot \mathcal{J}(x,E(t)).$$

where, by stationarity of μ, $\mathcal{J}$ is time independent. Supposing that h and $\mathbf{J}$ have natural symmetries under lattice translations and rotations we infer that $\mathcal{J}$ vanishes at constant E and then locality assumptions of the type we assumed for h imply

$$\mathcal{J}(x,E) = \sum_y \kappa(x,y,E)\nabla E(y).$$

Hence the annealed dynamics is a *discrete nonlinear diffusion*

$$E(t+1) - E(t) = \nabla \cdot \kappa(E(t)) \nabla E(t)$$

provided the diffusion matrix $\kappa(E(t))$ is positive.

Let now

$$\beta(t, x, E(t)) = J(t, x, E(t)) - \mathcal{J}(x, E(t))$$

be the fluctuating part. Then slow dynamics becomes

$$E(t+1) - E(t) = \nabla \cdot \kappa(E(t)) \nabla E(t) + \nabla \cdot \beta(t, E(t))$$

$$\mathbb{E}\ \beta(t, E) = 0$$

i.e. a *nonlinear diffusion* with *a random drift*. In a physical model one would expect $\kappa(E(t))$ to be positive although not necessarily uniformly in E. If furthermore β turned out to be a small perturbation quenched diffusion might be provable. In what follows we will make such assumptions and then indicate how to establish diffusion.

Before stating the assumptions let us make one more reduction. It is reasonable to assume $E = 0$ is preserved by the dynamics. This then implies $\beta(t, 0) = 0$. Let us study the linearization at $E = 0$:

$$E(t+1) - E(t) = \nabla \cdot \kappa(0) \nabla E(t) + \nabla \cdot (D\beta(0, t) E(t)) \tag{23}$$

or, in other words

$$E_x(t+1) = \sum_y p_{xy}(t) E_y(t) \tag{24}$$

with

$$\sum_x p_{xy}(t) = 1.$$

Since $E \geq 0$ we have $p_{xy} \geq 0$ i.e. $p_{xy}(t)$ are *transition probabilities of a random walk*. $p_{xy}(t)$ is space and time dependent and random i.e. it defines a *random walk in random environment*.

7. Random Walk in Nonlinear Random Environment

Consider a random walk defined by the transition probability matrix $p_{xy}(t)$ at time t. $p(t) = p(t, \omega)$ is taken random defined on some probability space Ω. We suppose the law of p is invariant under translations in space and time. Define

$$\|E\| := \sup_x |E(x)|(1 + |x|)^{d+a} \tag{25}$$

for some $a > 0$. Let, at $t = 0$, $\|E\| < \infty$. We say the walk defined by p is has a diffusive scaling limit if there exists C, κ such that *almost surely* in ω

$$\lim_{L\to\infty} \|L^d E(L^2 t, L\cdot) - Ct^{-d/2} E^*_\kappa(\cdot/\sqrt{t}\|) = 0 \tag{26}$$

where $E^*_\kappa(x) = e^{-x^2/4\kappa}$. In other words

$$L^d E(L^2 t, Lx) \sim Ct^{-d/2} e^{-x^2/4\kappa t}$$

as $L \to \infty$.

We prove this for a *non-linear perturbation* of RWRE. Let us state the assumptions for the random dynamical system eq. (22). We assume Φ is C^2 in $\|E\|_1 < \delta$ and satisfies

Positivity: $\Phi(E) \geq 0$ for $E \geq 0$.

Conservation law:

$$\sum_x \Phi(t, x, E) = \sum_x E_x$$

Weak nonlinearity:

$$\left| \frac{\partial^2 \Phi(t, x, E)}{\partial E_y \partial E_z} \right| \leq e^{-|x-y|-|x-z|}$$

Write the average map

$$\mathbb{E}\Phi(t, x, E) = \sum_y T(x - y)E_y + o(E).$$

Ellipticity: T generates a diffusive random walk on $\mathbb{Z}^d$.

Write

$$\Phi(t, x, E) - \mathbb{E}\Phi(t, x, E) := \nabla \cdot b(t, x, E).$$

Weak correlations. Assume

$$b(t, x, E) = \sum_{A \subset \mathbb{Z}^d \times [0,t]} b_A(t, x, E) \tag{27}$$

with

$$|b_A(t, x, E)| \leq \epsilon e^{-d((x,t)\cup A)}$$

and b_A, b_B are *independent* if $A \cap B = \emptyset$.

Remark. A representation of the form (27) arises from the model we have discussed above with the proviso that b_A, b_B are independent only in the case the θ dynamics is local, i.e. $h = 0$ in eq. (18). For the general h there is weak dependence that can be handled.

Theorem 7.1. *Under the above assumptions and δ, ϵ small enough the random dynamical system Φ_t is diffusive, almost surely in ω.*

8. Renormalization Group for Random Coupled Maps

The proof of Theorem 7.1. [9] is based on a renormalization group method introduced in [10] and [11]. Let us introduce the scaling transformation S_L:

$$(S_L E)(x) = L^d E(Lx). \tag{28}$$

where $L > 1$. Fix L and define, for each $n \in \mathbb{N}$, *renormalized energies*

$$E_n(t) = S_{L^n} E(L^{2n} t).$$

We can then rephrase the scaling limit (26) as

$$\lim_{n\to\infty} L^{nd} E(L^{2n}t, L^n x) = \lim_{n\to\infty} E_n(t, x).$$

$E_n(t)$ inherits dynamics from E. We will call this the *renormalized dynamics*:

$$E_n(t+1) = \Phi_n(t, E_n(t)).$$

Explicitely we have

$$\Phi_n(t) = S_{L^n}(\Phi(L^{2n}t + L^{2n} - 1) \circ \cdots \circ \Phi(L^{2n}t)) S_{L^{-n}}.$$

The dynamics changes with scale as

$$\Phi_{n+1} = \mathcal{R}\Phi_n$$

with

$$\mathcal{R}\Phi(t, \cdot) = S_L \Phi(t_{L^2}) \circ \cdots \circ \Phi(t_1) S_L^{-1}$$

with $t_1 = L^2 t$ and $t_{L^2} = L^2(t+1) - 1$.

$\mathcal{R}$ is the the *Renormalization group flow* in a space of random dynamical systems. We prove: almost surely the renormalized maps converge

$$\mathcal{R}^n f \to f^*$$

where the fixed point is *nonrandom and linear*:

$$f^*(E) = e^{\kappa\Delta} E.$$

Moreover, the renormalized energies converge almost surely to the fixed point

$$\left\| E_n(t, \cdot) - \frac{C}{t^{d/2}} E^*_\kappa(\cdot/\sqrt{t}) \right\| \to 0$$

which is the diffusive scaling limit.

These results may be summarized by saying that both the randomness and the nonlinearity are *irrelevant* in the RG sense. Let us finish by sketching the reasons for this.

We start by considering the linear problem

$$D\Phi(t,x,0)E = \sum_y p_{xy}(t)E_y.$$

Then $DR\Phi = p'$ with

$$p'(t)_{xy} = L^d(p(L^2(t+1)-1)\dots p(L^2t))_{LxLy}.$$

Write

$$p_{xy}(t) = T(x-y) + \nabla_y \cdot c_{xy}(t)$$

with $\mathbb{E}p = T$ and $\mathbb{E}c = 0$. Then, for $p' = T' + \nabla c'$ we get

$$T'(x-y) = L^d T^{L^2}(Lx - Ly) + r(x-y) \tag{29}$$

where r is an expectation of a polynomial in c. For the noise we get

$$\nabla_x c'_{xy} = L^d \sum_{t=1}^{L^2} \sum_{uv} T^t(Lx-u)\nabla_u c_{uv}(t) T^{L^2-t-1}(v-Ly) + \gamma_{xy}. \tag{30}$$

where γ involves quadratic and higher order polynomials in c.

Ignoring first r we get for the average flow

$$T_n = L^{nd} T^{L^{2n}}(L^n \cdot)$$

i.e.

$$\hat{T}_n(k) = \hat{T}^{L^{2n}}(k/L^n).$$

Write $\hat{T}(k) = 1 - ck^2 + o(k^2)$. Then as $n \to \infty$:

$$\hat{T}_n(k) \to e^{-ck^2}$$

explaining the fixed point.

Similarly, ignoring γ the noise is driven by the linear map

$$\mathcal{L}c_{xy}(0) = L^{d-1} \sum_{t=1}^{L^2} \sum_{uv} T^t(Lx-u) c_{uv}(t) T^{L^2-t-1}(v-Ly).$$

The variance of $\mathcal{L}c$ contracts:

$$\mathbb{E}(\mathcal{L}c)^2 \sim L^{-d} E c^2. \tag{31}$$

The intuitive reason behind this is the following. Take e.g. $x = y = 0$. For t of order L^2, $T^t(Lx-u) \sim L^{-d}e^{-|x-u/L|}$. Hence the u and the v sums are localized in an L cube at origin. Since $c_{uv}(t)$ has exponential decay in $|u-v|$

$$\mathcal{L}c_{00}(0) \sim L^{-d-1}\sum_{t=1}^{L^2}\sum_{|u|<L} c_{uu}(t). \tag{32}$$

Since correlations of c decay exponentially in space and time (32) is effectively a sum of L^{d+2} independent random variables of variance $L^{-2d-2}(\mathbb{E}c)^2$ thus leading to (31).

Taking into account the corrections r and γ in (29) and (30) we conclude that the variance contracts as

$$\mathbb{E}(c_n)^2 \sim \epsilon_n = L^{-nd}\epsilon.$$

The iteration of the mean becomes

$$T_{n+1} = L^d T_n^{L^2}(L\cdot) + \mathcal{O}(\epsilon_n). \tag{33}$$

The fixed point is the same but the $\mathcal{O}(\epsilon_n)$ renormalizes the diffusion constant κ at each iteration step (less and less as $n \to \infty$).

There is a problem however once we try to make this perturbative analysis rigorous. Deterministically the noise is *relevant*: from (32) we see that $\|\mathcal{L}c\|_\infty$ can be as big as $\mathcal{O}(L)\|c\|_\infty$. This means that there are unlikely events in the environment where the random walk develops a drift. We write

$$|c_n(t,,E)| \le L^{N_n(x)-bn}.$$

Then $N_n(x)$ can be (very) large, but with (very) small probability:

$$\mathrm{Prob}(N_n(x) > N) \le e^{-KN}$$

with K large.

Finally, to control the *nonlinear* contributions to Φ_n we show that the second derivative $D_E^2\Phi$ is irrelevant in all dimensions due to the scaling of E:

$$\mathcal{R}\Phi(t,x,E) = L^d(\Phi(t_{L^2}) \circ \cdots \circ \Phi(t_1))(Lx, L^{-d}E(\cdot/L)).$$

9. Towards Hamiltonian Systems

The coupled map lattices we have discussed are an alternative microscopic model with a local conservation law that under a macroscopic limit gives rise to diffusion. To be realistic they should however share some features with the Hamiltonian systems that are more familiar and physically relevant. From this point of view there is a lot missing from our analysis.

The first problem to understand is to go beyond the perturbative analysis around $E = 0$ (i.e. zero temperature). Then the equation (24) picks also a driving term.

The second unnatural assumption is the E-independence of the θ dynamics. In a realistic model rare configurations of E can slow down the θ dynamics. Also the annealed system is probably not uniformly elliptic as we assumed and the random drift can create traps in the environment with long lifetimes.

All these issues can and should be be studied with the renormalization group approach sketched above.

References

[1] H. Spohn, The phonon Boltzmann equation, properties and link to weakly anharmonic lattice dynamics, J. Stat. Phys. **124**, 1041–1104 (2006)

[2] J. Lukkarinen, H. Spohn, Weakly nonlinear Schrdinger equation with random initial data, arXiv:0901.3283v1 [math-ph]

[3] J.Bricmont, A. Kupiainen, Approach to equilibrium for the phonon Boltzmann equation Commun.Math. Phys. **281**, 179–202 (2008)

[4] J.Bricmont, A. Kupiainen, On the derivation of Fourier's law for coupled anharmonic oscillators, Commun.Math. Phys. **274**, 555–626 (2007)

[5] L. Bunimovich, C. Liverani, A. Pellegrinotti and Yu. Suhov, Ergodic systems of n balls in a billiard table, Commun.Math. Phys. **146**, 357 (1992)

[6] C.Mejia-Monasterio, H. Larralde, and F. Leyvraz. Coupled normal heat and matter transport in a simple model system. Phys. Rev. Lett. **86** (2001), 5417?5420.

[7] J.-P. Eckmann and L.-S. Young. Temperature profiles in Hamiltonian heat conduction. Europhysics Letters **68** (2004), 790–796.

[8] J.Bricmont, A. Kupiainen, High Temperature Expansions and Dynamical Systems, Commun.Math.Phys. **178**, 703–732 (1996)

[9] J.Bricmont, A. Kupiainen, In preparation

[10] J.Bricmont, A. Kupiainen, Random Walks in Asymmetric Random Environments, Comm.Math.Phys. **142**, 345–420 (1991)

[11] J.Bricmont, A. Kupiainen, Renormalization Group and Asymptotics of Solutions of Nonlinear Parabolic Equations, Commun. Pure.Appl.Math. **47**, 893–922 (1994)

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Noncommutative Geometry and Arithmetic

Matilde Marcolli*

Abstract

This is an overview of recent results aimed at developing a geometry of noncommutative tori with real multiplication, with the purpose of providing a parallel, for real quadratic fields, of the classical theory of elliptic curves with complex multiplication for imaginary quadratic fields. This talk concentrates on two main aspects: the relation of Stark numbers to the geometry of noncommutative tori with real multiplication, and the shadows of modular forms on the noncommutative boundary of modular curves, that is, the moduli space of noncommutative tori.

Mathematics Subject Classification (2010). 11M55.

Keywords. Noncommutative tori, real multiplication, Stark numbers, real quadratic fields, spectral triples, noncommutative boundary of modular curves, modular shadows, quantum statistical mechanics.

1. Introduction

The last few years have seen the development of a new line of investigation, aimed at applying methods of noncommutative geometry and theoretical physics to address questions in number theory. A broad overview of some of the main directions in which this area has progressed can be found in the recent monographs [41] and [14]. In this talk I am going to focus mostly on a particular, but in my opinion especially promising, aspect of this new and rapidly growing field, which did not get sufficient attention in [14], [41]: the question of developing an appropriate geometry underlying the abelian extensions of real quadratic fields. This line of investigation was initially proposed by Manin in [27], [28], as the "real multiplication program" and it aims at developing

*Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Mail Code 253-37, 1200 E.California Blvd, Pasadena, CA 91125, USA.
E-mail: matilde@caltech.edu.

within noncommutative geometry a parallel to the classical theory of elliptic curves with complex multiplication, and their role in the explicit construction of abelian extensions of imaginary quadratic fields, which would work for real quadratic fields. I am going to give an overview of the current state of the art in addressing this problem, by focusing on those aspects I have been more closely involved with.

There are two complementary approaches to developing a noncommutative geometry of real quadratic fields. One is based on working with noncommutative tori as substitutes for elliptic curves, focussing on those whose real parameter is a quadratic irrationality, which have non trivial self Morita equivalences, analogous to the complex multiplication phenomenon for elliptic curves. This approach requires constructing suitable functions on these spaces, which replace the coordinates of the torsion points of elliptic curves, hence the problem of finding suitable algebraic models for noncommutative tori. I will concentrate here especially on the question of how to express certain numbers, the Stark numbers, which conjecturally generate abelian extensions of real quadratic fields, in terms of the geometry of noncommutative tori.

The other complementary approach deals with a noncommutative space that parameterizes noncommutative tori up to Morita equivalence. This is sometimes referred to as the "invisible boundary" of the modular curves, since it parameterizes those degenerations of elliptic curves with level structure that are no longer expressible in algebro-geometric terms but that continue to exist as noncommutative tori. A related adelic version includes degenerations of the level structure and gives rise to a quantum statistical mechanical system based on the commensurability relation of lattices with possibly degenerate level structures, whose zero temperature equilibrium states, evaluated on an algebra of arithmetic elements should conjecturally provide generators of abelian extensions. The main problem in this approach is to obtain the right algebra of functions on this invisible boundary, which should consist of holographic images, or "shadows", that modular forms on the bulk space cast upon the invisible boundary.

2. Elliptic Curves and Noncommutative Tori

Elliptic curves are among the most widely studied objects in mathematics, whose pervasive presence in geometry, arithmetic and physics has made them a topic of nearly universal interest across mathematical disciplines. In number theory, one of the most famous manifestations of elliptic curves is through the theory of complex multiplication and the abelian class field theory problem (Hilbert 12th problem) in the case of imaginary quadratic fields.

The analytic model of an elliptic curve is the complex manifold realized as a quotient $E_\tau(\mathbb{C}) = \mathbb{C}^2/\Lambda$ with $\Lambda = \mathbb{Z} + \mathbb{Z}\tau$ or with the Jacobi uniformization $E_q(\mathbb{C}) = \mathbb{C}^*/q^{\mathbb{Z}}$ with $|q| < 1$. The endomorphism ring of an elliptic curves is a copy of $\mathbb{Z}$, except in the special case of elliptic curves with complex

multiplication where $\mathrm{End}(E_\tau) = \mathbb{Z} + fO_{\mathbb{K}}$, with $O_{\mathbb{K}}$ the ring of integers of an imaginary quadratic field and $f \geq 1$ an integer (the conductor).

A beautiful result in number theory relates the geometry of elliptic curves with complex multiplication to the explicit class field theory problem for imaginary quadratic fields: the explicit construction of generators of abelian extensions with the Galois action.

There are two formulations of this construction, one that works directly with the CM elliptic curves, and the coordinates of their torsion points, and one that works with the values of modular forms on the CM points in the moduli space of elliptic curves. (We refer the reader to [25], [49] for more information on this topic.)

As I will explain in the rest of the paper, both approaches have a noncommutative geometry analog in the case of real quadratic fields, which is in the process of being developed into a tool suitable for the investigation of the corresponding class field theory problem.

In the elliptic curve point of view, one knows that the maximal abelian extension $\mathbb{K}^{ab}$ of an imaginary quadratic field $\mathbb{K} = \mathbb{Q}(\sqrt{-d})$ has explicit generators

$$\mathbb{K}^{ab} = \mathbb{K}(t(E_{\mathbb{K},\mathrm{tors}}), j(E_{\mathbb{K}})),$$

where t is a coordinate on the quotient $E_{\mathbb{K}}/\mathrm{Aut}(E_{\mathbb{K}}) \simeq \mathbb{P}^1$ and $j(E_{\mathbb{K}})$ is the j-invariant.

I will explain below, based on a result of [37], how one can obtain an analog of the quotient $E_{\mathbb{K}}/\mathrm{Aut}(E_{\mathbb{K}})$ in the noncommutative geometry context for real quadratic fields. I will also mention some current approaches aimed at identifying the correct analog of the j-invariant in that setting.

Currently, the main problem in extending this approach to real quadratic fields via noncommutative geometry lies in the fact that, while elliptic curves have, besides the analytic model as quotients, an algebraic model as algebraic curves defined by polynomial equations, their noncommutative geometry analogs, the noncommutative tori, have a good analytic model, but not yet a fully satisfactory algebraic model. I will comment more on the current state of the art on this question in §3.2 below.

The other point of view, based on the moduli space, considers all elliptic curves, parameterized by the modular curve $X_\Gamma(\mathbb{C}) = \mathbb{H}/\Gamma$, with $\mathbb{H}$ the complex upper half plane and $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ acting on it by fractional linear transformations. One considers then the field F of modular functions. In this setting, the explicit class field theory result for imaginary quadratic fields is stated in terms of the generators

$$\mathbb{K}^{ab} = \mathbb{K}(f(\tau),\ f \in F,\ \tau \in \text{ CM points of } X_\Gamma),$$

the values of modular functions at CM points. The Galois action of $\mathrm{Gal}(\mathbb{K}^{ab}/\mathbb{K})$ is induced by the action of the automorphism group $\mathrm{Aut}(F)$ of the modular field.

The case of the explicit class field theory of $\mathbb{Q}$, the Kronecker–Weber theorem, can be formulated in terms of a special degenerate case of elliptic curves. When the parameter q in the elliptic curve $E_q(\mathbb{C})$ tends to a root of unity, or equivalently when the parameter $\tau \in \mathbb{H}$ tends to a rational point in the real line, the elliptic curve degenerates to a cylinder, the multiplicative group $\mathbb{C}^* = \mathbb{G}_m(\mathbb{C})$. The maximal abelian extension of $\mathbb{Q}$ is then generated by the torsion points of this degenerate elliptic curve,

$$\mathbb{Q}^{ab} = \mathbb{Q}(\mathbb{G}_{m,\mathrm{tors}}),$$

that is, by the roots of unity, the cyclotomic extensions.

The first case of number fields for which a solution to the explicit class field theory problem is not known is that of the real quadratic fields $\mathbb{K} = \mathbb{Q}(\sqrt{d})$. The approach currently being developed via noncommutative geometry is based on the idea of relating this case also to a special degenerate case of elliptic curves, the *noncommutative tori*. Manin's "Real multiplication program" [27], [28], to which I will return in the following, aims at building for noncommutative tori a parallel to the theory of complex multiplication for elliptic curves.

When the modulus q of the elliptic curve $E_q(\mathbb{C})$ tends to a point $\exp(2\pi i\theta)$ on the unit circle $S^1 \subset \mathbb{C}^*$ which is not a root of unity, or equivalently when $\tau \in \mathbb{H}$ tends to an irrational point on the real line, the elliptic curve degenerates in a much more drastic way. The action of $\mathbb{Z}$ by irrational rotations on the unit circle has dense orbits, so that the quotient, in the usual sense, does not deliver any interesting space that can be used to the purpose of doing geometry. This prevents one from considering such degenerations of elliptic curves in the usual algebro-geometric or complex-analytic world.

Noncommutative geometry, however, is explicitly designed in such a way as to treat "bad quotients" so that one can continue to make sense of ordinary geometry on them as if they were smooth objects. The main idea of how one does that is, instead of collapsing points via the equivalence relation of the quotient operation, one keeps all the identifications explicit in the groupoid describing the equivalence. More precisely, the algebra of functions on the quotient is replaced by a noncommutative algebra of functions on the graph of the equivalence relation with the associative convolution product dictated by the transitivity property of the equivalence relation,

$$(f_1 \star f_2)(x,y) = \sum_{x\sim z\sim y} f_1(x,z) f_2(z,y).$$

More precisely, in the case of the action of a discrete group G on a (compact) topological space X, the resulting algebra of (continuous) functions on the quotient is the *crossed product algebra* $C(X) \rtimes_\alpha G$, where the associative, noncommutative product is given by $(fU_g)(hU_{g'}) = f\alpha_g(h)U_{gg'}$, with $\alpha_g(h)(x) = h(g^{-1}(x))$.

In the case of the quotient of S^1 by the action of $\mathbb{Z}$ generated by $\exp(2\pi i\theta)$, an irrational rotation $\theta \in \mathbb{R} \smallsetminus \mathbb{Q}$, the quotient is therefore described by the

algebra $C(S^1) \rtimes_\theta \mathbb{Z}$. This is by definition the algebra $\mathcal{A}_\theta$ of continuous functions on the noncommutative torus $\mathbb{T}_\theta$ of modulus θ.

An equivalent description of the irrational rotation algebra $\mathcal{A}_\theta$ is as the universal C^*-algebra generated by two unitaries U, V with the commutation relation $VU = e^{2\pi i\theta}UV$. It has a smooth structure given by the smooth subalgebra of series $\sum_{n,m} a_{n,m}U^nV^m$ with rapidly decaying coefficients (*cf.* [8]).

Morita equivalence is the correct notion of isomorphism for noncommutative spaces, and it can be formulated in terms of the existence of a bimodule that implements an equivalence between the categories of modules for the two algebras. The algebras $\mathcal{A}_{\theta_1}$ and $\mathcal{A}_{\theta_2}$ are Morita equivalent if and only if there exists a $g \in \mathrm{SL}_2(\mathbb{Z})$ acting on $\mathbb{R}$ by fractional linear transformations, such that $\theta_1 = g\theta_2$, see [8], [48]. The bimodules realizing the Morita equivalences between noncommutative tori are obtained explicitly in [8] in terms of spaces of Schwartz functions on the line, and in [48] via a construction of projectors.

One can also describe the irrational rotation algebra of the noncommutative torus as a twisted group algebra $C^*(\mathbb{Z}^2, \sigma_\theta)$, with the cocyle

$$\sigma_\theta((n,m),(n',m')) = \exp(-2\pi i(\xi_1 nm' + \xi_2 mn')), \tag{2.1}$$

with $\theta = \xi_2 - \xi_1$. This is the norm closure of the action of the twisted group ring on $\ell^2(\mathbb{Z}^2)$ with the generators U and V are given by

$$Uf(n,m) = e^{-2\pi i\xi_2 n} f(n,m+1), \qquad Vf(n,m) = e^{-2\pi i\xi_1 m} f(n+1,m).$$

This description of the noncommutative torus is especially useful in the noncommutative geometry models of the integer quantum Hall effect, where this noncommutative space replaces the Brillouin zone in the presence of a magnetic field, see [3], [43].

3. *L*-functions, Solvmanifolds, and Noncommutative Tori

I give an overview here of recent progress in understanding the geometry of a special class of noncommutative tori, which have real multiplication, realized by nontrivial self Morita equivalences. These are the quantum tori $\mathbb{T}_\theta$ with $\theta \in \mathbb{R}$ a quadratic irrationality. In particular, I will focus on a result from [37] that realizes certain L-functions associated to real quadratic fields in terms of Riemannian and Loretzian geometry on the noncommutative tori with real multiplication.

3.1. Noncommutative tori with real multiplication. The starting observation of Manin's "Real multiplication program" is the following. The elliptic curves with complex multiplication are the only ones that have additional nontrivial endomorphisms, by the ring of integers $O_{\mathbb{K}}$ of an imaginary

quadratic field, and they correspond to lattices $\Lambda \subset \mathbb{C}$ that are $O_{\mathbb{K}}$-submodules with $\Lambda \otimes_{O_{\mathbb{K}}} \mathbb{K} \cong \mathbb{K}$, which corresponds to the parameter τ being a CM point of $\mathbb{H}$ for the imaginary quadratic field $\mathbb{K}$. In the same way, the noncommutative tori $\mathcal{A}_\theta$ for which the modulus θ is a real multiplication point in $\mathbb{R}$, in a real quadratic field $\mathbb{K} \subset \mathbb{R}$, have non-trivial self Morita equivalences, which play the role of the additional automorphisms of the CM elliptic curve.

3.2. Analytic versus algebraic model. A good part of the recent work on noncommutative tori with real multiplication was aimed at developing an algebraic model for these objects, in addition to the analytic model as quotients and crossed product algebras.

The most interesting approach to algebraic models for noncommutative tori is the one developed in [47], which is based on enriching the bimodules that give the self Morita equivalences with a "complex structure", in the sense of [16]. These are parameterized by the choice of an auxiliary elliptic curve E, or equivalently by a modulus $\tau \in \mathbb{H}$ up to $\mathrm{SL}_2(\mathbb{Z})$. By a suitable construction of morphisms, one obtains in this way a category of holomorphic vector bundles and a fully faithful functor to the derived category $D^b(E)$ of coherent sheaves on the auxiliary elliptic curve. The image is given by stable objects in the heart of a nonstandard t-structure, which depends on the parameter θ of the irrational rotation algebra $\mathcal{A}_\theta$ of the noncommutative torus. The real multiplication gives rise to autoequivalences of $D^b(E)$ preserving the t-structure.

This then makes it possible to associate to a noncommutative torus $\mathbb{T}_\theta$ with real multiplication a noncommutative algebraic variety, in the sense of [1]. These are described by graded algebras of the form

$$A_{F,O} = \bigoplus_{n \geq 0} \mathrm{Hom}(O, F^n(O)) \tag{3.1}$$

where O is an object of an additive category and F is an additive functor. In the case of the noncommutative tori of [47], the additive category is the heart of the t-structure in $D^b(E)$, the object O is $\mathcal{A}_\theta$, and F is induced by real multiplication, tensoring with the bimodule that generates the nontrivial self Morita equivalences.

The resulting ring was then related in [53] to the ring of quantum theta functions. These provide a good theory of theta functions for noncommutative tori developed in [29], [30]. As in the case of the classical theta functions, these can be constructed in terms of Heisenberg groups as a deformation of the classical case, see [29] (further elaborated upon in [46].) The relation between the quantum theta functions and the explicit construction of bimodules over noncommutative tori via projections was established in [4].

The arithmetic properties of the algebras of [47] were studied in [45], in terms of an explicit presentation of the twisted homogeneous coordinate rings (3.1) for real multiplication noncommutative tori, which involves modular forms of cusp type with level specified by an explicitly determined congruence subgroup.

A field of definition for these arithmetic structures on noncommutative tori can then be specified in terms of the field of definition of the auxiliary elliptic curve. It is not yet clear whether this approach to defining algebraic models for noncommutative tori with real multiplication can be successfully employed to provide a substitute for the coordinates of torsion points of elliptic curves in the CM case.

There is, however, another approach which works directly with the analytic model of noncommutative tori and with the candidate generators for abelian extensions of real quadratic fields given by Stark numbers.

3.3. Stark numbers and *L*-functions. There is in number theory a conjectural candidate for explicit generators of abelian extensions of real quadratic fields, in the form of Stark numbers, [51]. These are obtained by considering a family of L-functions associated to lattices $L \subset \mathbb{K}$ in a real quadratic field. In the notation of [27], one considers an $\ell_0 \in O_{\mathbb{K}}$, with the property that the ideals $\mathfrak{b} = (L, \ell_0)$ and $\mathfrak{a} = (\ell_0)\mathfrak{b}^{-1}$ are coprime with $\mathfrak{f} = L\mathfrak{b}^{-1}$. Let U_L denote the set of units of $\mathbb{K}$ such that $u(\ell_0 + L) = \ell_0 + L$, and let $'$ denote the Galois conjugate, with $N(\ell) = \ell\ell'$. One then considers the function

$$\zeta(L, \ell_0, s) = \mathrm{sign}(\ell_0')\ N(\mathfrak{b})^s \sum_{\ell \in (\ell_0 + L)/U_L} \frac{\mathrm{sign}(\ell')}{|N(\ell)|^s}. \tag{3.2}$$

The associated Stark number is then

$$S_0(L, \ell_0) = \exp\left(\frac{d}{ds} \zeta(L, \ell_0, s)|_{s=0} \right). \tag{3.3}$$

Part of the "real multiplication program" of [27], [28] is the question of providing an interpretation of these numbers directly in terms of the geometry of noncommutative tori with real multiplication.

To understand how one can relate these numbers to RM noncommutative tori and to a suitable noncommutative space that plays the role of the quotient $E/\mathrm{Aut}(E)$ of a CM elliptic curve, we concentrate here on the case of a closely related L-function, the Shimizu L-function of a lattice in a real quadratic field.

The lattice $L \subset \mathbb{K}$ define a lattice $\Lambda = \iota(L) \subset \mathbb{R}^2$ via the two embeddings $L \ni \ell \mapsto (\ell, \ell')$. The group V of units of $\mathbb{K}$ satisfying

$$V = \{u \in O_{\mathbb{K}}^* \,|\, uL \subset L,\ \iota(u) \in (\mathbb{R}_+^*)^2\}$$

has generator a unit ϵ and it acts on Λ by $(x, y) \mapsto (\epsilon x, \epsilon' y)$. The Shimizu L-function is then given by

$$L(\Lambda, s) = \sum_{\mu \in (\Lambda \smallsetminus \{0\})/V} \frac{\mathrm{sign}(N(\mu))}{|N(\mu)|^s}. \tag{3.4}$$

This corresponds to the case $\ell_0 = 0$ of (3.2), with the sum avoiding the point $0 \in \Lambda$.

3.4. Solvmanifolds and noncommutative spaces. The hint on how the L-function (3.4) is related to RM noncommutative tori comes from a well known result of Atiyah–Donnelly–Singer [2], which proved a conjecture of Hirzebruch relating the Shimizu L-function to the signature of the Hilbert modular surfaces, through the computation of the eta invariant of a 3-dimensional solvmanifold which is the link of an isolated singularity of the Hilbert modular surface. The result of [2] is in fact more generally formulated for Hilbert modular varieties and L-functions of totally real fields, but for our purposes we concentrate on the real quadratic case only.

Although it does not look like it at first sight, and it was certainly not formulated in those terms, the result of [2] is in fact saying something very useful about the geometry of noncommutative tori with real multiplication, as I explained in [37].

A first observation is the fact that, in noncommutative geometry, one often has a way to construct a commutative model, up to homotopy, of a noncommutative space describing a bad quotient. The idea is similar to the use of homotopy quotients in topology, and is closely related to the Baum–Connes conjecture. In fact, the latter can be seen as the statement that invariants of noncommutative spaces, such as K-theory, can be computed geometrically using a commutative model as homotopy quotient.

As we recalled above, a "bad quotient" can be described by a noncommutative space with algebra of functions an associative convolution algebra, the crossed product algebra in the case of a group action. In particular, we consider the noncommutative space describing the quotient $\mathbb{T}_\theta/\mathrm{Aut}(\mathbb{T}_\theta)$ of a noncommutative torus with real multiplication by the automorphisms coming from the group V of units in the real quadratic field $\mathbb{K}$ preserving the lattice $L \subset \mathbb{K}$. The quotient of the action of the group of units V on the noncommutative torus with real multiplication is described by the crossed product algebra $\mathcal{A}_\theta \rtimes V$. This can also be described by a twisted group algebra of the form

$$\mathcal{A}_\theta \rtimes V \cong C^*(\mathbb{Z}^2 \rtimes_{\varphi_\epsilon} \mathbb{Z}, \tilde{\sigma}_\theta), \tag{3.5}$$

where, after identifying the lattice Λ with $\mathbb{Z}^2$ on a given basis, the action of the generator ϵ of V on Λ is implemented by a matrix $\varphi_\epsilon \in \mathrm{SL}_2(\mathbb{Z})$, and one correspondingly identifies the semidirect product $S(\Lambda, V) = \Lambda \rtimes_\epsilon V$ with $\mathbb{Z}^2 \rtimes_{\varphi_\epsilon} \mathbb{Z}$. The cocycle $\tilde{\sigma}$ is given by

$$\tilde{\sigma}_\theta((n,m,k),(n',m',k')) = \sigma_\theta((n,m),(n',m')\varphi_\epsilon^k). \tag{3.6}$$

This is indeed a cocycle for $S(\Lambda, V)$, for $\xi_2 = -\xi_1 = \theta/2$, since in this case (2.1) satisfies $\sigma((n,m)\gamma,(n',m')\gamma) = \sigma((n,m),(n',m'))$, for $\gamma \in \mathrm{SL}_2(\mathbb{Z})$.

Groups of the form $S(\Lambda, V)$ satisfy the Baum–Connes conjecture. This implies that the quotient noncommutative space $\mathbb{T}_\theta/\mathrm{Aut}(\mathbb{T}_\theta)$, with algebra of coordinates $C^*(S(\Lambda, V), \tilde{\sigma}_\theta)$, admits a good homotopy quotient model. In this case, as shown in [37], this homotopy quotient can be identified explicitly as the 3-dimensional smooth solvmanifold X_ϵ obtained as the quotient of $\mathbb{R}^2 \rtimes_\epsilon \mathbb{R}$

by the group $S(\Lambda, V)$. This is the same 3-manifold that gives the link of the singularity in the Hilbert modular surface in [2], whose eta invariant computes the signature defect.

Another way to describe this 3-dimensional solvmanifold, with its natural metric, is in terms of Hecke lifts of geodesics to the space of lattices (see [27] and [37]). For $t \in \mathbb{R}$ one considers the lattice in $\mathbb{R}^2$ of the form

$$\iota_t(L) = \{(xe^t, ye^{-t}) \,|\, (x,y) \in \Lambda\},$$

with $\iota_1(L) = \Lambda$, as above. Then one has a fibration $T^2 \to S(\Lambda, V) \to S^1$, where the base S^1 is a circle of length $\log \epsilon$, identified with the closed geodesic in $X_\Gamma(\mathbb{C})$ corresponding to the geodesic in $\mathbb{H}$ with endpoints $\theta, \theta' \in \mathbb{R}$, for $\{1, \theta\}$ a basis of the real quadratic field $\mathbb{K}$. The fiber over $t \in S^1$ is the 2-torus $T^2_t = \mathbb{R}^2/\iota_t(L)$.

The result of [2] can then be reintepreted as saying that the spectral theory of the Dirac operator on the 3-dimensional solvmanifold X_ϵ can be decomposed into a contribution coming from the underlying noncommutative space $\mathbb{T}_\theta/\mathrm{Aut}(\mathbb{T}_\theta)$, and an additional spurious part, which depends on the choice of a homotopy model for this quotient. The part coming from the underlying noncommutative torus is the one that recovers the Shimizu L-function and that is responsible for the signature defect computed in [2].

3.5. Spectral triples. To understand how the Dirac operator on the manifold X_ϵ can be related to a Dirac operator on the noncommutative space, one can resort to the general formalism of *spectral triples* in noncommutative geometry [9]. One encodes metric geometry on a noncommutative space by means of the data $(\mathcal{A}, \mathcal{H}, D)$ of a representation on a Hilbert space $\mathcal{H}$ of a dense subalgebra $\mathcal{A}$ of the algebra of coordinates, together with a self-adjoint (unbounded) operator D on $\mathcal{H}$ with compact resolvent, satisfying the condition that commutators $[D, a]$ with elements of the algebra are bounded operators. This plays the role of an abstract Dirac operator which provides the metric structure.

3.6. The Shimizu *L*-function and noncommutative tori. One can then relate the Dirac operator on X_ϵ to a spectral triple on the noncommutative torus with real multiplication, which recovers the Shimizu L-function, in two steps, [37]. The first makes use of the isospectral deformations of manifolds introduced in [12]. Given a smooth spin Riemannian manifold X, which admits an action of a torus T^2 by isometries, one can construct a deformation of X to a family of noncommutative spaces X_η, parameterized by a real parameter $\eta \in \mathbb{R}$, with algebras of coordinates $\mathcal{A}_{X_\eta}$, in such a way that, if $(C^\infty(X), L^2(X,S), D)$ is the original spectral triple describing the ordinary spin geometry on X, then the data $(\mathcal{A}_{X_\eta}, L^2(X,S), D)$ still give a spectral triple on X_η. In this way, one can isospectrally deform the fibration $T^2 \to X_\epsilon \to S^1$ to a noncommutative space $X_{\epsilon,\theta}$, which is a fibration $\mathbb{T}_\theta \to X_{\epsilon,\theta} \to S^1$, where

$\mathbb{T}_\theta$ is the noncommutative torus with real multiplication. One then checks that, up to a unitary equivalence, the restriction of the Dirac operator to the fiber $\mathbb{T}_\theta$ gives a spectral triple on this noncommutative torus with Dirac operator of the form

$$D_{\theta,\theta'} = \begin{pmatrix} 0 & \delta_{\theta'} - i\delta_\theta \\ \delta_{\theta'} + i\delta_\theta & 0 \end{pmatrix}, \tag{3.7}$$

with $\{1, \theta\}$ the basis for the real quadratic field $\mathbb{K}$ and θ' the Galois conjugate of θ. The derivations δ_θ and $\delta_{\theta'}$ act as

$$\delta_\theta \psi_{n,m} = (n + m\theta)\psi_{n,m}, \quad \text{and} \quad \delta_\theta \psi_{n,m} = (n + m\theta')\psi_{n,m},$$

and they correspond to leafwise derivations $e^t\partial_x$ and $e^{-t}\partial_y$ on the tori T^2_t. The Dirac operator $D_{\theta,\theta'}$ decomposes into a product of an operator with spectrum $\mathrm{sign}(N(\mu))|N(\mu)|^{1/2}$, which recovers the Shimizu L-function, and a term whose spectrum only depends on the powers ϵ^k on the unit ϵ, see §7 of [37].

3.7. Lorentzian geometry. An important problem in the context of noncommutative geometry is extending the formalism of spectral triples from Riemannian to Lorentzian geometries. This is especially important in the particle physics and cosmology models based on spectral triples and the spectral action functionals, see [7], [44]. A proposal for Lorentzian noncommutative geometries, based on Krein spaces replacing Hilbert spaces in the indefinite signature context, was developed in [52].

Another interesting aspect of the geometry of noncommutative tori with real multiplication is the fact that the spectral triples described above admit a continuation to a Lorentzian geometry, based on considering the norm of the real quadratic field $N(\lambda) = \lambda_1\lambda_2 = (n+m\theta)(n+m\theta')$ as the analog of the wave operator in momentum space, with modes $\Box_\lambda = N(\lambda)$. The Krein involution is constructed using the Galois conjugation of the real quadratic field, and the Wick rotation to Euclidean signature of the resulting Lorentzian Dirac operator $\mathcal{D}_\mathbb{K}$ on $\mathbb{T}_\theta$, with $\mathcal{D}^2_{\mathbb{K},\lambda} = \Box_\lambda$, recovers the Dirac operator $D_{\theta,\theta'}$. The eta function of the Lorentzian spectral triple is a product

$$\eta_{\mathcal{D}_\mathbb{K}}(s) = L(\Lambda, V, s)Z(\epsilon, s),$$

of the Shimizu L-function and a function that only depends on the unit ϵ.

3.8. Quantum field theory and noncommutative tori. This result of [37] recalled above explains how certain number theoretic L-functions associated to real quadratic fields, such as the Shimizu L-function or, more generally, the zeta functions of (3.2) arise from the noncommutative geometry of noncommutative tori with real multiplication $\mathbb{T}_\theta$ and their quotients $\mathbb{T}_\theta/\mathrm{Aut}(\mathbb{T}_\theta)$.

One would then like to explain the meaning in terms of noncommutative geometry of numbers of the form $\exp(L'(0))$, where $L(s)$ is one of these L-functions, since this is the class of numbers that the Stark conjectures propose

as conjectural generators of abelian extensions. While there is at present no completely satisfactory answer to this second question, I describe here some work in progress in which I am trying to provide such interpretation in terms of quantum field theory.

It should not come as a surprise that one would aim at realizing numbers of arithmetic significance in terms of quantum field theory. In fact, there is a broad range of results (see [40] for an overview) relating the evaluation of Feynman integrals in quantum field theory to the arithmetic geometry of motives.

Here the point of connection is the zeta function regularization method in quantum field theory. This expresses the functional integral that gives the partition function as

$$\int e^{-\langle \phi, D\phi \rangle} \mathcal{D}[\phi] \sim (\det(D))^{-1/2},$$

where the quantity $\det(D)$ here is obtained through the zeta function regularization, using the zeta function $\zeta_D(s) = \mathrm{Tr}(|D|^{-s})$ of the operator D and setting $\det(D) = \exp(-\zeta'_D(0))$.

To adapt this to the setting described above of spectral triples on a noncommutative torus with real multiplication, one can use the fact that there is a well developed method [21] for doing quantum field theory on finite projective modules, that is, for fields that are sections of "bundles over noncommutative spaces". This formalism was developed completely explicitly in [21] for the case of finite projective modules over noncommutative tori. In the case with real multiplication, one has a preferred choice of a QFT, namely the one associated to the bimodule that generates the non-trivial self Morita equivalences that give the RM structure. A description of the numbers (3.3) in terms of this quantum field theory is work in progress [42].

4. The Noncommutative Boundary of Modular Curves

In the case of the imaginary quadratic fields, as we mentioned above, the other approach to constructing abelian extensions is by considering, instead of individual CM elliptic curves, the CM points on the moduli space of elliptic curves.

In terms of noncommutative tori, one can similarly consider a moduli space that parameterizes the equivalence classes under Morita equivalence. This itself is described by a noncommutative space, which corresponds to the quotient of $\mathbb{P}^1(\mathbb{R})$ by the action of $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ by fractional linear transformations. As a noncommutative space, this is described by the crossed product algebra $C(\mathbb{P}^1(\mathbb{R})) \rtimes \Gamma$. This space parameterizes degenerate lattices where $\tau \in \mathbb{H}$ becomes a point $\theta \in \mathbb{R}$. One thinks of this space as the "invisible boundary" of the modular curve $X_\Gamma(\mathbb{C})$. It complements the usual boundary $\mathbb{P}^1(\mathbb{Q})/\Gamma$ (the cusp point corresponding to the degeneration of the elliptic curve $E_\tau(\mathbb{C})$ to the

multiplicative group $\mathbb{G}_m(\mathbb{C})$) with the irrational points $(\mathbb{R} \smallsetminus \mathbb{Q})/\Gamma$, treated as a noncommutative space. These irrational points account for the degenerations to noncommutative tori, "invisible" to the usual world of algebraic geometry but nonetheless existing as noncommutative spaces.

In this approach, the main question becomes identifying what remnants of modularity one can have on this "invisible boundary" and what replaces evaluating a modular form at a CM point in this setting.

4.1. Modular shadow play. A phenomenon similar to the "holography principle" (also known as AdS/CFT correspondence) of string theory relates the noncommutative geometry of the invisible boundary of the modular curves to the algebraic geometry of the classical "bulk space" $X_\Gamma(\mathbb{C})$ (see [36]). For example, it was shown in [34] that the K-theory of the crossed product algebra $C(\mathbb{P}^1(\mathbb{R})) \rtimes \Gamma$ recovers Manin's modular complex [31], which gives an explicit presentation of the homology of the modular curves $X_\Gamma(\mathbb{C})$.

A way of inducing on the noncommutative boundary $\mathbb{P}^1(\mathbb{R})/\Gamma$ a class of functions corresponding to modular forms on the bulk space $X_\Gamma(\mathbb{C})$ was given in [34], [35] in terms of a Lévy–Mellin transform, which can be thought of as creating a "holographic image" of a modular form on the boundary.

Consider a complex valued function f which is defined on pairs (q, q') of coprime integers $q \geq q' \geq 1$, satisfying $f(q, q') = O(q^{-\epsilon})$ for some $\epsilon > 0$. For $x \in (0, 1]$ set

$$\ell(f)(x) = \sum_{n=1}^{\infty} f(q_n(x), q_{n-1}(x)),$$

where the $q_n(x)$ are successive denominators of the continued fraction expansion of x. Lévy's lemma (see [34]) shows that one has

$$\int_0^1 \ell(f)(x)dx = \sum_{q \geq q' \geq 1; (q,q')=1} \frac{f(q, q')}{q(q + q')}.$$

This identity can be used to recast identities of modular forms in terms of integrals on the boundary $\mathbb{P}^1(\mathbb{R})$. For example, it is shown in [34] that one can use the function

$$f(q, q') = \frac{q + q'}{q^{1+t}} \{0, q'/q\}$$

with $\Re(t) > 0$ and $\{0, q'/q\}$ the classical modular symbol, together with the identity of [31],

$$\sum_{d|m} \sum_{b=1}^{d} \int_{\{0, b/d\}} \omega = (\sigma(m) - c_m) \int_0^{i\infty} \pi_\Gamma^*(\omega),$$

where $\pi_\Gamma^*(\omega)/dz$ is a cusp form for $\Gamma = \Gamma_0(p)$, with p a prime, which is an eigenvector of the Hecke operator T_m with eigenvalue c_m, with $p \not| m$, and with

$\sigma(m)$ the sum of the divisors of m. One then obtains an identity of the form

$$\int_0^1 dx \sum_{n=0}^{\infty} \frac{q_{n+1}(x)+q_n(x)}{q_{n+1}(x)^{1+t}} \int_{\{0,q_n(x)/q_{n+1}(x)\}} \omega = \left(\frac{\zeta(1+t)}{\zeta(2+t)} - \frac{L_\omega^{(p)}(2+t)}{\zeta^{(p)}(2+t)^2} \right) \int_0^{i\infty} \pi^*(\omega),$$

where $L_\omega^{(p)}$ and $\zeta^{(p)}$ are the Mellin transform and zeta function with omitted p-th Euler factor. Other such examples were given in [34], [38].

This type of identities, recasting integrals of cusp forms on modular symbols in terms of integrals along the invisible boundary of a transform of the modular form producing a function on the boundary, can be formulated more generally and more abstractly as a way of obtaining "shadows" of modular forms on the boundary. In [35] one considers pseudomeasures associated to pair of rational points on the boundary with values in an abelian group, $\mu : \mathbb{P}^1(\mathbb{Q}) \times \mathbb{P}^1(\mathbb{Q}) \to W$, satisfying $\mu(x,x) = 0$, $\mu(x,y) + \mu(y,x) = 0$, and $\mu(x,y) + \mu(y,z) + \mu(z,x) = 0$. In particular the modular pseudomeasures satisfy $\mu\gamma(x,y) = \gamma\mu(x,y)$, or an analogous identity twisted by a character, where $\gamma(x,y) = (\gamma(x), \gamma(y))$ is the action of a finite index $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{P}^1(\mathbb{Q})$ by fractional linear transformations. The classical Hecke operators act on modular pseudomeasures. Pseudomeasures can be equivalently formulated in terms of currents on the tree $\mathcal{T}$ of $\mathrm{PSL}_2(\mathbb{Z})$ embedded in the hyperbolic plane $\mathbb{H}$. In terms of noncommutative spaces, they can also be described as group homomorphisms $\mu : K_0(C(\partial\mathcal{T}) \rtimes \Gamma) \to W$.

Integration along geodesics in $\mathbb{H}$ of holomorphic functions vanishing at cusps define pseudomeasures on the boundary. It is shown in [35] that one can obtain "shadows" of modular symbols on the boundary by the following procedure. Given a finite index subgroup $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$ and a weight $w \in \mathbb{N}$, let $\mathcal{S}_{w+2}(\Gamma)$ be the $\mathbb{C}$-vector space of cusp forms $f(z)$ of weight $w+2$ for Γ, holomorphic on $\mathbb{H}$ and vanishing at cusps. Let $\mathcal{P}_w$ be the space of homogeneous polynomials of degree w in two variables and let W be the space of linear functionals on $\mathcal{S}_{w+2} \otimes \mathcal{P}_w$. Then

$$\mu(x,y) : f \otimes P \mapsto \int_x^y f(z)P(z,1)dz$$

defines a W-valued modular pseudomeasure, which is the shadow of the higher weight modular symbol of [50].

A general formulation is the given in [35], which encompasses the averaging techniques over successive convergents of the continued fraction expansion, used in [34] to relate Mellin transforms of weight-two cusp forms to quantities defined entirely on the noncommutative boundary of the modular curves. One considers a class of functions $\ell(f)(x) = \sum_I f(I)\chi_I(x)$ that are formal infinite linear combinations of characteristic functions of "primitive intervals" in $[0,1]$, with

coefficients $f(I)$ in an abelian group. More generally, this may depend on an additional regularization parameter, $\ell(f)(x,s)$. The primitive intervals are those of the form $I=(g(\infty),g(0))$ with $g\in \mathrm{GL}_2(\mathbb{Z})$. Pseudomeasures are completely determined by their values on these intervals. The Lévy–Mellin transform is then defined in [35] as $\mathcal{LM}(s)=\int_0^{1/2}\ell(f)(x,s)dx$. The integration over $[0,1/2]$ instead of $[0,1]$ keeps symmetry into account. When applied to a pseudomeasure obtained as the shadow of a modular symbol, for an $\mathrm{SL}_2(\mathbb{Z})$-cusp form this gives back the usual Mellin transform.

The formalism of pseudomeasures was also used in [33] to describe modular symbols for Maass wave forms, based on the work of Lewis–Zagier [26]. In particular, Manin gives in [33] an interpretation of the Lévy–Mellin transform of [35] as an analog at arithmetic infinity (at the archimedan prime) of the p-adic Mellin–Mazur transform.

4.2. Modular shadows and the Kronecker limit formula. Modular pseudomeasures with values in a Γ-module W, with $\Gamma=\mathrm{PSL}_2(\mathbb{Z})$, give rise to 1-cocycles, by setting $\phi^\mu_x(\gamma)=\mu(\gamma x,x)$. The cocycle condition $\phi(\gamma_1\gamma_2)=\phi(\gamma_1)+\gamma_1\phi(\gamma_2)$ follows from the modularity of μ together with the relations $\mu(x,x)=0$, $\mu(x,y)+\mu(y,x)=0$, and $\mu(x,y)+\mu(y,z)+\mu(z,x)=0$, see [35]. Conversely, any cocycle with $\phi(\sigma)=\phi(\tau)$, where σ and τ are the generators of order two and three of $\Gamma=\mathbb{Z}/2\mathbb{Z}\star\mathbb{Z}/3\mathbb{Z}$. In fact, a pseudomeasure is determined by the relations $(1+\sigma)\mu(0,\infty)=0$ and $(1+\tau+\tau^2)\mu(0,\infty)=0$, while a 1-cocycle is determined by the relations $(1+\sigma)\phi(\sigma)=0$ and $(1+\tau+\tau^2)\phi(\tau)=0$.

An interesting recent result [54] gives a construction of a modular pseudomeasure involved in a higher Kronecker limit formula for real quadratic fields. The pseudomeasure takes values in $C(\mathbb{P}^1(\mathbb{R}))$ with the action of Γ of weight $2k$. One considers a function

$$\psi_{2k}(x)=\mathrm{sign}(x)\sum_{p,q\geq 0}^{*}(p|x|+q)^k,$$

where the $*$ on the sum means that the sum is for $(p,q)\neq(0,0)$ and that the terms with $p=0$ or $q=0$ are counted with a coefficient $1/2$. The modular pseudomeasure is given by setting $\mu(0,\infty)=\psi_{2k}$, since $\psi_{2k}=\phi(\sigma)=\phi(\tau)$ determines a 1-cocycle. For $x>0$ the function $\psi_{2k}(x)$ is also expressed in terms of the derivatives of functions $\mathcal{F}_{2k}$, constructed in terms of the digamma function $\Gamma'(x)/\Gamma(x)$, which give the higher Kronecker limit formula proved in [54] as

$$\zeta(\mathfrak{b},k)=\sum_{Q\in\mathrm{Red}(\mathfrak{b})}(\mathcal{D}_{k-1}\mathcal{F}_{2k})(Q),$$

where $\zeta(\mathfrak{b},s)=\sum_{\mathfrak{n}\in\mathfrak{b}}N(\mathfrak{n})^{-s}$ and $\mathrm{Red}(\mathfrak{b})$ is the set of reduced quadratic forms in the class $\mathfrak{b}$, by seeing narrow ideal classes as Γ-orbits on the set of integer quadratic forms. The $\mathcal{D}_{k-1}$ are differential operators of order k mapping differentiable functions of one variable to functions of two variables, given explicitly in [54].

In particular, as shown in [54], one can use this higher Kronecker limit formula to evaluate Stark numbers, as values at $k = 1$ of the zeta-functions rather than as derivatives at zero. This provides an alternative way of connecting Stark numbers to the geometry of noncommutative tori, not by working with a single noncommutative torus with real multiplication, but with their noncommutative moduli space and the modular shadows.

4.3. Quantum modular forms. There is at present another approach to extending modularity to the boundary, in a form that arises frequently in very different contexts, such as quantum invariants of 3-manifolds. Zagier recently developed [55] a notion of *quantum modular forms*, which encompasses all these phenomena. The idea is that, instead of the usual properties of modular forms, namely a holomorphic function on $\mathbb{H}$ satisfying the modularity property

$$(f|_k\gamma)(z) := f\left(\frac{az+b}{cz+d}\right)(cz+d)^{-k} = f(z),$$

one has a function f defined on $\mathbb{P}^1(\mathbb{Q})$, for which the function

$$h_\gamma(x) = f(x) - (f|_k\gamma)(x), \tag{4.1}$$

which measures the failure of modularity, extends to a continuous or even (piecewise) analytic function on $\mathbb{P}^1(\mathbb{R})$.

A more refined notion of "strong" quantum modular form prescribes that, besides having evaluations at all rational points, the function f also has a formal Taylor series expansion at all $x \in \mathbb{Q}$, and (4.1) is an identity of formal power series. Typical examples of strong quantum modular forms described in [55] have the additional property that the function $f : \mathbb{P}^1(\mathbb{Q}) \to \mathbb{C}$ extends to a function $f : (\mathbb{C} \smallsetminus \mathbb{R}) \cup \mathbb{Q} \to \mathbb{C}$, which is analytic on $\mathbb{C} \smallsetminus \mathbb{R}$, and whose asymptotic expansion approaching a point $x \in \mathbb{Q}$ along vertical lines agrees with the formal Taylor series of f at x. Such quantum modular forms can be thought of as two analytic functions, on the upper and lower half plane, respectively, that communicate across the rational points on the boundary.

There are two observations one can make to relate this setting to noncommutative geometry. One is that, in the case of quantum modular forms, one is dealing with functions f defined on the rational points of the boundary, while the "invisible boundary" consisting of the irrational points is seen only through the associated function h_γ which measures the failure of modularity of f. Thus, the object that should be interpreted in terms of the noncommutative space $C(\mathbb{P}^1(\mathbb{R})) \rtimes \Gamma$ is the h_γ rather than the quantum modular form f itself.

Another observation is that a similar setting, with functions that have evaluations and Taylor expansions at all rational points, is provided by the Habiro ring of "analytic functions of roots of unity" [23]. This was, in fact, also developed to deal with the same phenomenology of quantum invariants of 3-manifolds, such as the Witten–Reshetikhin–Turaev invariants, which typically have a value at each root of unity as well as a formal Taylor expansion, the

Ohtsuki series. Those strong quantum modular forms that satisfy an additional integrality condition needed in the construction of the Habiro ring may be thought of as objects satisfying a partial modularity property (through the associated h_γ) among these analytic functions of roots of unity. Several significant examples of quantum modular forms given in [55] indeed define elements in the Habiro ring.

The functions in the Habiro ring were recently interpreted in [32] as providing the right class of functions to do analytic geometry over the "field with one element" $\mathbb{F}_1$. This was then reformulated in the setting of noncommutative geometry in [39] using the notion of endomotives developed in [10] (see also §4 of [14]) which are a category of noncommutative spaces combining Artin motives with semigroup actions, together with the relation between the endomotive associated to abelian extensions of $\mathbb{Q}$ and Soulé's notion of geometry over $\mathbb{F}_1$, established in [11]. The same noncommutative space and some natural multivariable generalizations are related in [39] to another notion of geometry over $\mathbb{F}_1$ developed by Borger [5] in terms of consistent lifts of Frobenius encoded in the structure of a Λ-ring.

5. Quantum Statistical Mechanics and Number Fields

The description of the boundary of modular curves in terms of the noncommutative space $C(\mathbb{P}^1(\mathbb{R})) \rtimes \Gamma$, for Γ a finite index subgroup of the modular group, accounts for degenerations of lattices with level structures, to degenerate lattices (pseudolattices in the terminology of [27]). In the adelic description, this would correspond to degenerating lattices with level structures at the archimedean component. In fact, one can also consider degenerating the level structures at the non-archimedean components. This leads to another noncommutative space, which contains the usual modular curves, and which also contains in its compactification the invisible boundary described above.

In [13] such a noncommutative space of adelic degenerations of lattices with level structures was described as the moduli space of 2-dimensional $\mathbb{Q}$-lattices up to commensurability and up to a scaling relation. A $\mathbb{Q}$-lattice is a pair of a lattice Λ together with a group homomorphism $\phi : \mathbb{Q}^2/\mathbb{Z}^2 \to \mathbb{Q}\Lambda/\Lambda$ which is a possibly degenerate level structure (it is not required to be an isomorphism). Commensurability means that $\mathbb{Q}\Lambda_1 = \mathbb{Q}\Lambda_2$ and $\phi_1 = \phi_2$ modulo $\Lambda_1 + \Lambda_2$. The scaling is by an action of $\mathbb{C}^*$. The corresponding noncommutative space is the convolution algebra of functions $f((\Lambda,\phi),(\Lambda',\phi'))$ of pairs of commensurable lattices that are of degree zero for the $\mathbb{C}^*$-action, with the convolution product

$$(f_1 \star f_2)((\Lambda,\phi),(\Lambda',\phi')) = \sum_{(\Lambda'',\phi'')\sim(\Lambda,\phi)} f_1((\Lambda,\phi),(\Lambda'',\phi''))f_2((\Lambda'',\phi''),(\Lambda',\phi')).$$

This admits a convenient parameterization in terms of coordinates (g, ρ, z) with $g \in \mathrm{GL}_2^+(\mathbb{Q})$, $\rho \in M_2(\hat{\mathbb{Z}})$, and $z \in \mathbb{H}$.

The advantage of adopting this point of view is that the resulting noncommutative space, whose algebra of coordinates I denote here by $\mathcal{A}_{GL(2),\mathbb{Q}}$, has a natural time evolution, by the covolume of lattices

$$\sigma_t(f)((\Lambda,\phi),(\Lambda',\phi')) = \left(\frac{\mathrm{covol}(\Lambda')}{\mathrm{covol}(\Lambda)}\right)^{it} f((\Lambda,\phi),(\Lambda',\phi')).$$

5.1. Zero temperature states and modular forms. The extremal low temperature KMS equilibrium states for the dynamical system $(\mathcal{A}_{GL(2),\mathbb{Q}}, \sigma)$ are parameterized by those $\mathbb{Q}$-lattices for which ϕ is an isomorphism (the invertible ones). Thus the set of extremal low temperature KMS states can be identified ([13], [14] §3) with the usual Shimura variety $\mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})/\mathbb{C}^*$. This can be thought of as the set of the classical points of the noncommutative space $\mathcal{A}_{GL(2),\mathbb{Q}}$.

The adelic group $\mathbb{Q}^*\backslash\mathrm{GL}_2(\mathbb{A}_{\mathbb{Q},f})$ acts as symmetries of this quantum statistical mechanical system, with the subgroup $\mathrm{GL}_2(\hat{\mathbb{Z}})$ of $\mathrm{GL}_2(\mathbb{A}_{\mathbb{Q},f}) = \mathrm{GL}_2^+(\mathbb{Q}) \cdot \mathrm{GL}_2(\hat{\mathbb{Z}})$ acting by automorphisms, and $\mathrm{GL}_2^+(\mathbb{Q})$ by endomorphisms, and the quotient by $\mathbb{Q}^*$ eliminating the inner symmetries that act trivially on the KMS states.

The zero temperature extremal KMS states, defined in [13] as weak limits of the positive temperature ones, have the property that, when evaluated at elements of a $\mathbb{Q}$-algebra $\mathcal{M}_{GL(2),\mathbb{Q}}$ of unbounded multipliers of $\mathcal{A}_{GL(2),\mathbb{Q}}$, they give values that are evaluations of modular forms $f \in F$ at points in $\mathbb{H}$. Under the identification $\mathbb{Q}^*\backslash\mathrm{GL}_2(\mathbb{A}_{\mathbb{Q},f}) \cong \mathrm{Aut}(F)$, for a generic set of points $\tau \in \mathbb{H}$ the action of symmetries of the dynamical system is intertwined with the action of automorphisms of the modular field. This is very much like the $GL(1)$-case of [6], which corresponds, in the same setting, to the case of 1-dimensional $\mathbb{Q}$-lattices.

5.2. Imaginary quadratic fields. One can recast in this setting of quantum statistical mechanical systems the case of imaginary quadratic fields, [15]. One considers a similar convolution algebra for 1-dimensional $\mathbb{K}$-lattices, for $\mathbb{K} = \mathbb{Q}(\sqrt{-d})$ and realizes it as a subalgebra $\mathcal{A}_\mathbb{K}$ of the algebra of commensurability classes of 2-dimensional $\mathbb{Q}$-lattices recalled above. In this case, the extremal low temperature KMS states are parameterized by the invertible $\mathbb{K}$-lattices, which are labelled by a CM point in $\mathbb{H}$ and an element in $\hat{O}_\mathbb{K}$. The evaluation of extremal zero temperature KMS states on the restriction of the algebra $\mathcal{M}_{GL(2),\mathbb{Q}}$ to $\mathcal{A}_\mathbb{K}$ then give evaluations of modular forms at CM points and the action of symmetries induces the correct action of $\mathrm{Gal}(\mathbb{K}^{ab}/\mathbb{K})$.

5.3. Quantum statistical mechanical systems for number fields. The construction of [15] of quantum statistical mechanical systems

$(\mathcal{A}_{\mathbb{K}}, \sigma)$ associated to imaginary quadratic fields, using the system for 2-dimensional $\mathbb{Q}$-lattices of [13], was generalized in [22] to a construction of a similar system for an arbitrary number field, using a generalization of the $GL(2)$-system to quantum statistical mechanical systems associated to arbitrary Shimura varieties. Rewritten in the notation of [24] these quantum statistical mechanical systems $(\mathcal{A}_{\mathbb{K}}, \sigma)$ for number fields are given by semigroup crossed product algebras of the form

$$\mathcal{A}_{\mathbb{K}} = C(G_{\mathbb{K}}^{ab} \times_{\hat{O}_{\mathbb{K}}^*} \hat{O}_{\mathbb{K}}) \rtimes J_{\mathbb{K}}^+,$$

where $J_{\mathbb{K}}^+$ is the semigroup of integral ideals and $G_{\mathbb{K}}^{ab} = \mathrm{Gal}(\mathbb{K}^{ab}/\mathbb{K})$. These also admit an interpretation as convolution algebras of commensurability classes of 1-dimensional $\mathbb{K}$-lattices, see [24]. The time evolution is by the norm of ideals

$$\sigma_t(f) = f, \quad \forall f \in C(G_{\mathbb{K}}^{ab} \times_{\hat{O}_{\mathbb{K}}^*} \hat{O}_{\mathbb{K}}), \qquad \text{and} \qquad \sigma_t(\mu_{\mathfrak{n}}) = N(\mathfrak{n})^{it} \mu_{\mathfrak{n}}, \quad \forall \mathfrak{n} \in J_{\mathbb{K}}^+.$$

An explicit presentation for the algebras $\mathcal{A}_{\mathbb{K}}$ was obtained in [20], by embedding them into larger crossed product algebras. What is still missing in this general construction is the "algebra of arithmetic elements" replacing $\mathcal{M}_{GL(2),\mathbb{Q}}$, on which to evaluate the zero temperature extremal KMS states to get candidate generators of abelian extensions. In the particular case of the real quadratic fields, such an algebra would contain the correct replacement for the modular functions on the invisible boundary of the modular curves.

5.4. Noncommutative geometry and anabelian geometry.

The quantum statistical mechanical systems for number fields described above are explicitly designed to carry information on the abelian extensions of the field, hence they involve the abelianization of the absolute Galois group. However, it appears that these noncommutative spaces may in fact contain also the full "anabelian" geometry of number fields. This is presently being investigated in my joint work with Cornelissen [19]. The question is to what extent one can reconstruct the number field from the system $(\mathcal{A}_{\mathbb{K}}, \sigma)$. The fact that the partition function of this quantum statistical mechanical system is the Dedekind zeta function and that the evaluation of low temperature KMS states on elements in the algebra can be written in terms of Dirichlet series, shows that at least the system recovers the arithmetic equivalence class of the field. A similar results should in fact hold for function fields, where a version of these quantum statistical mechanical systems in the positive characteristic setting with partition function the Goss zeta function was developed in [17] (see [18] for the role of the Goss zeta function for arithmetic equivalence.) It is more subtle to see whether the system $(\mathcal{A}_{\mathbb{K}}, \sigma)$ recovers not only the field up to arithmetic equivalence but also up to isomorphism, [19].

References

[1] M. Artin, M. van der Bergh, *Twisted homogeneous coordinate rings*, J. Algebra 133 (1990) 249–271.

[2] M.F. Atiyah, H. Donnelly, I.M. Singer, *Eta invariants, signature defects of cusps, and values of L-functions*, Annals of Mathematics 118 (1983) 131–177.

[3] J. Bellissard, *Noncommutative geometry and quantum Hall effect*. Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), 1238–1246, Birkhäuser, 1995.

[4] F.P. Boca, *Projections in rotation algebras and theta functions*, Commun. Math. Phys. 202 (1999) 325–357.

[5] J. Borger, *Lambda-rings and the field with one element*, arXiv:0906.3146.

[6] J.B. Bost, A. Connes, *Hecke algebras, type III factors and phase transitions with spontaneous symmetry breaking in number theory*. Selecta Math. (N.S.) Vol.1 (1995) N.3, 411–457.

[7] A. Chamseddine, A. Connes, M. Marcolli, *Gravity and the standard model with neutrino mixing*, Advances in Theoretical and Mathematical Physics, 11 (2007) 991–1090.

[8] A. Connes, *C^* algèbres et géométrie differentielle*. C.R. Acad. Sci. Paris, Ser. A-B, 290 (1980) 599–604.

[9] A. Connes, *Geometry from the spectral point of view*. Lett. Math. Phys. 34 (1995), no. 3, 203–238.

[10] A. Connes, C. Consani, M. Marcolli, *Noncommutative geometry and motives: the thermodynamics of endomotives*, Advances in Math. 214 (2) (2007), 761–831.

[11] A. Connes, C. Consani, M. Marcolli, *Fun with $\mathbb{F}_1$*, J. Number Theory, Vol.129 (2009), N.6, 1532–1561.

[12] A. Connes, G. Landi, *Noncommutative manifolds, the instanton algebra and isospectral deformations*. Comm. Math. Phys. 221 (2001), no. 1, 141–159.

[13] A. Connes, M. Marcolli, *From physics to number theory via noncommutative geometry, Part I: Quantum statistical mechanics of $\mathbb{Q}$-lattices*, in Frontiers in Number Theory, Physics and Geometry, I pp.269–350, Springer Verlag, 2006.

[14] A. Connes, M. Marcolli, *Noncommutative geometry, quantum fields and motives*, Colloquium Publications, Vol.55, AMS, 2008.

[15] A. Connes, M. Marcolli, N. Ramachandran, *KMS states and complex multiplication*, Selecta Math. (N.S.) Vol.11 (2005) N.3–4, 325–347.

[16] A. Connes, M. Rieffel, *Yang–Mills for noncommutative two-tori*. in "Operator algebras and mathematical physics", pp.237–266, Contemp. Math., 62, AMS, 1987.

[17] C. Consani, M. Marcolli, *Quantum statistical mechanics over function fields*, Journal of Number Theory **123** (2) (2007) 487–528.

[18] G. Cornelissen, A. Kontogeorgis, L. van der Zalm, *Arithmetic equivalence, the Goss zeta function, and a generalisation*, preprint arxiv:0906.4424, to appear in J. Number Theory, 2009.

[19] G. Cornelissen, M. Marcolli, *Quantum statistical mechanics, L-series and anabelian geometry*, in preparation.

[20] J. Cuntz, X. Li, *C^*-algebras associated with integral domains and crossed products by actions on adele spaces*, arXiv:0906.4903.

[21] V. Gayral, J.H. Jureit, T. Krajewski, R. Wulkenhaar, *Quantum field theory on projective modules*. J. Noncommut. Geom. 1 (2007), no. 4, 431–496.

[22] E. Ha, F. Paugam, *Bost-Connes-Marcolli systems for Shimura varieties. I. Denitions and formal analytic properties*, IMRP Int. Math. Res. Pap. (2005), no. 5, 237–286.

[23] K. Habiro, *Cyclotomic completions of polynomial rings*, Publ. RIMS Kyoto Univ. (2004) Vol.40, 1127–1146.

[24] M. Laca, N.S. Larsen, S. Neshveyev, *On Bost-Connes types systems for number fields*, J. Number Theory 129 (2009), no. 2, 325–338.

[25] S. Lang, *Elliptic Functions*, (Second Edition), Graduate Texts in Mathematics, Vol.112, Springer Verlag, 1987.

[26] J. Lewis, D. Zagier, *Period functions for Maass wave forms, I*, Ann. of Math. (2) 153 (2001) N.1, 191–258.

[27] Yu.I. Manin, *Real multiplication and noncommutative geometry (ein Alterstraum)*. The legacy of Niels Henrik Abel, 685–727, Springer, Berlin, 2004.

[28] Yu.I. Manin, *Von Zahlen und Figuren*, in "Géométrie au XXe siècle. Histoire et horizons", pp.24–44, Hermann, 2005.

[29] Yu.I. Manin, *Theta functions, quantum tori and Heisenberg groups*. Lett. Math. Phys. 56 (2001), no. 3, 295–320.

[30] Yu.I. Manin, *Quantized theta-functions*. in "Common trends in mathematics and quantum field theories" (Kyoto, 1990). Progr. Theoret. Phys. Suppl. No. 102 (1990), 219–228 (1991).

[31] Yu.I. Manin, *Parabolic points and zeta functions of modular curves*, in "Selected Papers", pp.268–290, World Scientific, 1996.

[32] Yu.I. Manin, *Cyclotomy and analytic geometry over F_1*, arXiv:0809.1564.

[33] Yu.I. Manin, *Remarks on modular symbols for Maass wave forms*, arXiv:0803.3270.

[34] Yu.I. Manin, M. Marcolli, *Continued fractions, modular symbols, and noncommutative geometry*, Selecta Mathematica (New Series) Vol.8 N.3 (2002) 475–520.

[35] Yu.I. Manin, M. Marcolli, *Modular shadows and the Levy-Mellin infinity-adic transform*, in "Modular forms on Schiermonnikoog", Cambridge University Press, 2008, pp. 189–238.

[36] Yu.I. Manin, M. Marcolli, *Holography principle and arithmetic of algebraic curves*, Advances in Theoretical and Mathematical Physics, Vol.5, N.3 (2001) 617–650.

[37] M. Marcolli, *Solvmanifolds and noncommutative tori with real multiplication*, Communications in Number Theory and Physics, Volume 2, No. 2 (2008) 423–479.

[38] M. Marcolli, *Limiting modular symbols and the Lyapunov spectrum*, Journal of Number Theory, Vol.98 N.2 (2003) 348–376.

[39] M. Marcolli, *Cyclotomy and endomotives*, P-Adic Numbers, Ultrametric Analysis and Applications, Vol.1 (2009) N.3, 217–263.

[40] M. Marcolli, *Feynman motives*, World Scientific, 2009.

[41] M. Marcolli, *Arithmetic noncommutative geometry*, University Lectures Series, Vol.35, AMS, 2005.

[42] M. Marcolli, *Quantum field theory on noncommutative tori with real multiplication*, in preparation.

[43] M. Marcolli, V. Mathai, *Towards the fractional quantum Hall effect, a noncommutative geometry perspective*, in "Noncommutative Geometry and Number Theory", pp.235–262. Vieweg Verlag, 2006.

[44] M. Marcolli, E. Pierpaoli, *Early Universe models from noncommutative geometry*, arXiv:0908.3683.

[45] J. Plazas, *Arithmetic structures on noncommutative tori with real multiplication.* Int. Math. Res. Not. IMRN 2008, no. 2, Art. ID rnm147, 41 pp.

[46] J. Plazas, *Heisenberg modules over real multiplication noncommutative tori and related algebraic structures*, arXiv:0712.0279.

[47] A. Polishchuk, *Noncommutative two-tori with real multiplication as noncommutative projective varieties*, J. Geom. Phys. 50 (2004), no. 1–4, 162–187.

[48] M.A. Rieffel, C^**-algebras associated to irrational rotations*, Pacific J. Math. 93 (1981) 415–429.

[49] G. Shimura, *Arithmetic theory of automorphic functions*, Princeton, 1971.

[50] V. Shokurov, *Shimura integrals of cusp forms*, Math. USSR Izvestiya, Vol.16 (1981) N.3, 603–646.

[51] H.M. Stark, *L-functions at* $s = 1$*. IV. First derivatives at* $s = 0$, Adv. Math. 35 (1980) 197–235.

[52] A. Strohmaier, *On noncommutative and pseudo-Riemannian geometry.* J. Geom. Phys. 56 (2006), no. 2, 175–195.

[53] M. Vlasenko, *The graded ring of quantum theta functions for noncommutative torus with real multiplication.* Int. Math. Res. Not. 2006, Art. ID 15825, 19 pp.

[54] M. Vlasenko, D. Zagier, *Higher Kronecker "limit" formulas for real quadratic fields*, preprint 2010.

[55] D. Zagier, *Quantum modular forms*, preprint 2009.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Universality, Phase Transitions and Extended Scaling Relations

Vieri Mastropietro*

Abstract

The universality hypothesis in statistical physics says that a number of macroscopic critical properties are largely independent of the microscopic structure, at least inside a universality class of systems. In the case of planar interacting Ising models, like Vertex or Ashkin-Teller models, this hypothesis means that the critical exponents, though model dependent, verify a set of universal extended scaling relations. The proof of several of such relations has been recently achieved; it is valid for generic non solvable models and it is based on the Renormalization Group methods developed in the context of constructive Quantum Field Theory. Extensions to quantum systems and several challenging open problems will be also presented.

Mathematics Subject Classification (2010). Primary 82B20, 82B27, 82B28; Secondary 81T16, 81T17

Keywords. Universality, lattice Ising systems, critical phenomena, Renormalization Group, nonperturbative renormalization.

1. Phase Transitions and Critical Phenomena

The aim of Statistical mechanics is to predict the macroscopic properties of the matter starting from its microscopic atomic description, and it is still well explained by the words of Democritus (460-370 BC): "From the ordering and positions of the atoms the changes of the matter can be explained". In particular, one aims at understanding the phenomenon of *phase transitions*, in which a material modifies its macroscopic state while its microscopic components remain the same.

*The author thanks G. Benfatto, P.Falco and A. Giuliani, as several results reported in this paper were obtained with them.

Dipartimento di Matematica, Università di Roma "Tor Vergata", 00133 Roma, Italy.
E-mail: mastropi@axp.mat.uniroma2.it.

A phenomenon widely studied in statistical physics is magnetism. Several magnetic materials, at temperatures T lower than a certain critical temperature T_c, have a spontaneous magnetization which corresponds to a microscopically ordered phase; the system undergoes a phase transition at $h = 0$ changing suddenly from positive to negative magnetization reversing the magnetic field h. On the other hand, for $T \geq T_c$, there is no spontaneous magnetization (disordered phase); the point $h = 0, T = T_c$ is called *critical* and the properties close to it are particularly remarkable. In order to understand such a phenomenon we can describe the magnet in terms of *spin models*. We can imagine the magnet as made up of molecules sitting on the sites of a finite square lattice $\Lambda \subset \mathbb{Z}^d$, where Λ is centered around the origin and contains $|\Lambda| = L^d$ lattice sites. Each molecule can be regarded as a dipole pointing along a preferred axis, with two possible directions. The molecule at the point $\mathbf{x} \in \Lambda$ has then two possible configurations, which can be labelled by a spin variable $\sigma_{\mathbf{x}}$ with values 1 or -1. The "configurations" of the system consist of a set $\sigma = (\sigma_{\mathbf{x}_1}, ..., \sigma_{\mathbf{x}_{|\Lambda|}})$ of $|\Lambda|$ numbers such that $\sigma_{\mathbf{x}} = \pm$. The number of these configurations is $2^{|\Lambda|}$ and to each spin configuration a certain energy $H(\sigma)$ is assigned.

The postulates of statistical mechanics allow to compute macroscopic quantities appearing in the thermodynamical theory of the system starting from the microscopic energy $H(\sigma)$; indeed the *partition function* is

$$Z = \sum_{\sigma} e^{-\beta H(\sigma)} \tag{1.1}$$

where $\beta = (\kappa T)^{-1}$, T is the temperature and κ is the Boltzmann constant. The *free energy* per site is defined as $f_{\Lambda,\beta} = -\beta^{-1}|\Lambda|^{-1} \log Z$; it can be proved under rather general conditions on the energy, see *e.g.* Theorem 2.4.1. of [50], that the limit

$$f_\beta = -\beta^{-1} \lim_{|\Lambda|\to\infty} \frac{1}{|\Lambda|} \log Z \tag{1.2}$$

exists and is convex. Phase transitions, which can be present only in the *thermodynamic limit* $|\Lambda| \to \infty$, appear as non-analyticity points of f_β. The derivatives of f_β correspond to physical observables; for instance the *specific heat* is defined (when exists) as

$$C_v = -\frac{\partial}{\partial T} T^2 \frac{\partial}{\partial T} \frac{f_\beta}{T} \tag{1.3}$$

If O is some observable property of the system with value $O(\sigma)$ in the spin configuration σ, than its observed average thermodynamic value is

$$\langle O \rangle_{\Lambda,\beta} = \sum_{\sigma} O(\sigma) \frac{e^{-\beta H}}{Z} \tag{1.4}$$

and if $O_{\mathbf{x}}$ is a local monomial in the spin variables, like $\sigma_{\mathbf{x}}$ or $\sigma_{\mathbf{x}}\sigma_{\mathbf{x}'}$ (where $\mathbf{x}'$ is a nearest-neighbor of $\mathbf{x}$), its truncated correlations are

$$\lim_{\Lambda\to\infty} [\langle O_{\mathbf{x}} O_{\mathbf{y}} \rangle_{\Lambda,\beta} - \langle O_{\mathbf{x}} \rangle_{\Lambda,\beta} \langle O_{\mathbf{y}} \rangle_{\Lambda,\beta}] \tag{1.5}$$

The correlations measure, roughly speaking, the influence in a point $\mathbf{y}$ of a perturbation located at a certain point $\mathbf{x}$. The basic problem in equilibrium statistical physics is the computation of the free energy and its derivatives and of the other thermodynamic functions or correlations.

A particularly simple form for the energy H is the one in the *nearest-neighbor Ising model*

$$H_J(\sigma) = -J \sum_{<i,j>} \sigma_{\mathbf{x}_i}\sigma_{\mathbf{x}_j} - h\sum_{\mathbf{x}} \sigma_{\mathbf{x}} \tag{1.6}$$

where $J > 0$, $\sum_{<i,j>}$ means that the sum is over pairs $(\mathbf{x}_i, \mathbf{x}_j)$ of nearest neighbor of Λ. The model is an oversimplified description of a magnet with just one easy axis of magnetization; the first term in (1.6) takes into account the exchange energy between the dipoles (the contribution to the energy of two neighbor dipoles has opposite sign depending if the dipoles point in the same or in the opposite direction) and the second term takes into account the interaction with an external magnetic field.

In one dimension the model is solvable but the free energy is analytic in the thermodynamic limit; there are no phase transitions. In three dimension, no exact solution has been found; it is therefore particularly remarkable that, in two dimensions and with no magnetic field, the nearest-neighbor Ising model can be exactly solved and it exhibits a phase transition, that is a non analyticity point in the free energy as a function of temperature. The solution is due to Onsager [46]; before such result it was not even clear if the formalism of statistical mechanics can describe phase transition. The free energy can be computed (see §5 of [44] or [52]) and one can determine the critical temperature at which the thermodynamic functions have singularities; for instance the specific heat has a logarithmic divergence [1]

$$C_v \sim -\log|\beta - \beta_c| \tag{1.7}$$

where

$$\tanh(\beta_c J) = \sqrt{2} - 1 \tag{1.8}$$

and β_c is the *critical temperature*. Even if the model is so simplified with respect to a real magnet, it explains several of its physical properties; for instance, the solution predict a *spontaneous magnetization* which is vanishing at the critical temperature with a critical exponent. In general the critical behavior of the 2D Ising model (1.6) close to the critical point $h = 0, T = T_c$ is characterized by a set of J-independent *critical exponents*. For instance the *energy correlation function* $G_\beta(\mathbf{x} - \mathbf{y})$, defined by (1.5) with $O_{\mathbf{x}}$ chosen as $\sigma_{\mathbf{x}}\sigma_{\mathbf{x}'}$ ($\mathbf{x}'$ a nearest neighbor point of $\mathbf{x}$), decays at large distances for $\beta \neq \beta_c$ faster than any power of $\xi^{-1}|\mathbf{x} - \mathbf{y}|$, with ξ the *correlation length* diverging at $\beta = \beta_c$

$$\xi^{-1} \sim |\beta - \beta_c|^{\nu} \tag{1.9}$$

[1] we say that $X \sim Y$ if there are two constants c_1, c_2 such that $c_1 Y \leq X \leq c_2 Y$

with $\nu = 1$; moreover at $\beta = \beta_c$ the energy correlations decay with a power law

$$G_{\beta_c}(\mathbf{x} - \mathbf{y}) \sim \frac{1}{|\mathbf{x} - \mathbf{y}|^X} \tag{1.10}$$

with exponent $X = 2$. In general, the correlation length of the correlations at the critical point becomes infinite; a perturbation in a point have an influence at a very large distance from it.

2. Universality

The Ising model is an oversimplified description of a real system; in general, realistic models for matter are extremely complex and depend on a number of microscopic details; the computation of the physical observables, to be compared with experiments, is essentially hopeless. In this context, the *universality* hypothesis plays a crucial role; it says that the critical properties should be *insensitive* to the details of the microscopic description, at least inside a certain *universality class* of systems. By such hypothesis highly oversimplified models can be used to get information on realistic and complex systems close to criticality; this is a crucial property for having quantitative predictions.

In the case of systems in the class of universality of the Ising model, universality simply says that the *critical exponents* are the same. This is well confirmed in experiments: for instance the experimental value of the exponents of Carbon dioxide or Xenon coincide with the three dimensional Ising model ones (obtained by numerical simulations), see [27], even if, of course, the Hamiltonians describing such compounds is completely different. A spectacular confirmation of universality came from a recent experiment in a space mission [37], in which measured exponents for the λ-transition of Helium are coinciding with several digits with the ones of the three dimensional XY model. From the mathematical point of view, universality in dimensions equal to 4 and above is a consequence of a strengthened version of the central limit theorem, see [18] or [2]; in lower dimension the phenomenon is more subtle.

We will be interested from now on about the issue of universality in *two dimensions*. A basic question is to understand what happens if we consider, instead of (1.6), a more general model; namely if Λ is a square subset of $\mathbb{Z}^2$ of side L, and $\mathbf{x} = (x_0, x) \in \Lambda$

$$H(\sigma) = H_J(\sigma) + \lambda V(\sigma) \tag{2.1}$$

where H_J is the nearest neighbor Ising model (1.6), which can be written as

$$H_J(\sigma) = -J \sum_{j=0,1} \sum_{\mathbf{x} \in \Lambda} \sigma_{\mathbf{x}} \sigma_{\mathbf{x}+\mathbf{e}_j} \tag{2.2}$$

where $\mathbf{e}_0 = (0,1), \mathbf{e}_1 = (1,0)$, λ is a coupling and V is quartic in the spins; an example is

$$V(\sigma) = \sum_j \sum_{\mathbf{x},\mathbf{y}\in\Lambda} v(\mathbf{x}-\mathbf{y})\sigma_{\mathbf{x}}\sigma_{\mathbf{x}+\mathbf{e}_j}\sigma_{\mathbf{y}}\sigma_{\mathbf{y}+\mathbf{e}_j} \tag{2.3}$$

with

$$|v(\mathbf{x}-\mathbf{y})| \le e^{-\kappa_0|\mathbf{x}-\mathbf{y}|} \tag{2.4}$$

and κ_0 a constant; also a *next to nearest neighbor* interaction can be written in the form (2.3)

$$V(\sigma) = \sum_j \sum_{\mathbf{x}\in\Lambda} \sigma_{\mathbf{x}}\sigma_{\mathbf{x}+\mathbf{e}_j}\sigma_{\mathbf{x}+\mathbf{e}_j}\sigma_{\mathbf{x}+2\mathbf{e}_j} = \sum_j \sum_{\mathbf{x}\in\Lambda} \sigma_{\mathbf{x}}\sigma_{\mathbf{x}+2\mathbf{e}_j} \tag{2.5}$$

The Hamiltonian (2.1) is "physically equivalent" to (1.6), as from a physical point of view there is no reason for which only nearest-neighbor spins should interact; it is much more reasonable to assume that the interaction becomes weaker and weaker as more distant spins are considered. In the same way, it is also very natural to include interactions involving four or a greater number of spins. Even if the Hamiltonian (2.1) is physically equivalent to (1.6), an exact solution for it is not known and there is no exact way to compute the exponents.

It is generally believed that the Hamiltonian (2.1) is in the universality class of the 2D Ising model. Only very recently a universality result for (2.1) has been proved by Pinson and Spencer [57, 54] using as a starting point the Grassmann integral representation of the correlations (see below). They proved that for λ small enough the model (2.1) is critical at

$$\tanh \beta_c J = \sqrt{2} - 1 + O(\lambda) \tag{2.6}$$

and that the specific heat has the same logarithmic singularity (1.7) as in the nearest neighbor case. They have also shown that if $\beta \neq \beta_c$ the energy correlation decays at large distances faster than any power of $\xi^{-1}|\mathbf{x}-\mathbf{y}|$, with $\xi^{-1} \sim |\beta - \beta_c|$; moreover at $\beta = \beta_c$ it decays with a power law with exponent 2. The critical temperature depends an the coupling λ in (2.1) but the exponents are λ-independent.

There are however systems in which the exponents are not pure numbers but depend on all the microscopic structure; this happens in physical systems like planar magnetic materials, carbon nanotubes or spin chains like $KCuF_3$. In such cases universality acquires a more subtle form; it *does not mean*, as for models in the Ising class, that the exponents do not depend from the microscopic details (on the contrary they do); rather, it means that *there exist universal and model-independent relations allowing to express, for instance, all the exponents in terms of a few of them*. Even if the critical exponents depend on the extraordinarily complex microscopic details, the universal relations allow concrete and testable predictions in terms of a few measurable parameters.

The simplest class of models showing exponents depending on the Hamiltonian parameters is obtained by considering *two* planar Ising models coupled by a quartic interaction; the Hamiltonian is

$$H(\sigma,\sigma') = H_J(\sigma) + H_{J'}(\sigma') - \lambda V(\sigma,\sigma') \tag{2.7}$$

with H_J given by (2.2), V is a short ranged, quartic interaction in the spin and invariant under the spin exchange; an example is

$$V(\sigma) = \sum_{j=0,1} \sum_{\mathbf{x},\mathbf{y}\in\Lambda} v(\mathbf{x}-\mathbf{y})\sigma_{\mathbf{x}}\sigma_{\mathbf{x}+\mathbf{e}_j}\sigma'_{\mathbf{y}}\sigma'_{\mathbf{y}+\mathbf{e}_j} \tag{2.8}$$

with $v(\mathbf{x})$ a short range potential, and another example is provided by (2.12) below. We will be interested in particular in the specific heat C_v and the energy ($\epsilon = +$) and cross-over ($\epsilon = -$) correlations, defined as

$$G^{\epsilon}_{\beta}(\mathbf{x}-\mathbf{y}) = \lim_{\Lambda\to\infty} [\langle O^{\epsilon}_{\mathbf{x}} O^{\epsilon}_{\mathbf{y}}\rangle_{\Lambda,\beta} - \langle O^{\epsilon}_{\mathbf{x}}\rangle_{\Lambda,\beta}\langle O^{\epsilon}_{\mathbf{y}}\rangle_{\Lambda,\beta}] \qquad \epsilon = \pm \tag{2.9}$$

where

$$O^{\epsilon}_{\mathbf{x}} = \sum_{j=0,1} \sigma_{\mathbf{x}}\sigma_{\mathbf{x}+\mathbf{e}_j} + \epsilon \sum_{j=0,1} \sigma'_{\mathbf{x}}\sigma'_{\mathbf{x}+\mathbf{e}_j} \tag{2.10}$$

The model (2.7) describes *two* interacting magnetic layers, each of them described by an Ising model. Moreover several systems in statistical mechanics, like the *Ashkin-Teller* or the *Eight Vertex* models, can be rewritten as coupled Ising models, see [5].

In the Ashkin-Teller model the spin has *four* values A, B, C, D, and two neighbor spins are associated an energy E_1 for AA, BB, CC, DD, E_2 for AB, CD, E_3 for AC, BD, E_4 for AD, BC. It was proposed to describe the properties of certain alloys and it can be also seen as a more realistic generalization of the Ising model, as the assumption that the dipole can point only in two directions is rather crude; in real systems the dipole can point in any direction. It is easy to see that the Ashkin-Teller Hamiltonian can be written in the form (2.7) with a suitable choice of λ, J, J' and

$$V(\sigma) = \sum_{j=0,1} \sum_{\mathbf{x}\in\Lambda} \sigma_{\mathbf{x}}\sigma_{\mathbf{x}+\mathbf{e}_j}\sigma'_{\mathbf{x}}\sigma'_{\mathbf{x}+\mathbf{e}_j} \tag{2.11}$$

For a choice of parameters such that $J = J'$ the Ashkin-Teller model is called *isotropic*, while for $J \neq J'$ it is called *anisotropic*. When $\lambda = 0$ the model is exactly solvable as its Hamiltonian is the sum of two independent Ising models, and two *critical temperatures* are present which if $J \neq J'$ which reduce to one in the $J = J'$ case; no solution is known for $\lambda \neq 0$.

The Eight Vertex model is a generalization of the Vertex models introduced to describe the idrogen bounding in ice [5], and it can also be mapped in (2.7) with a suitable identification of the parameters; in such a case $J = J'$ and

$$V(\sigma) = \sum_{j=0,1} \sum_{\mathbf{x}\in\Lambda} \sigma_{\mathbf{x}+j(\mathbf{e}_0+\mathbf{e}_1)}\sigma_{\mathbf{x}+\mathbf{e}_0}\sigma'_{\mathbf{x}+j(\mathbf{e}_0+\mathbf{e}_1)}\sigma'_{\mathbf{x}+\mathbf{e}_1} \tag{2.12}$$

The exact solution for the Eight Vertex model was found by Baxter [4] in the early seventies. For a particular choice of the parameters the Eight Vertex model reduces to Six Vertex models, previously solved by Lieb [36] and Sutherland [56].

From the Baxter solution the specific heat α and the correlation length exponent ν can be computed (see (10.12.22), (10.12.23) of [5]) and it is found that they are *non constant functions of* λ and different from the Ising ones. This was considered somewhat surprising at the time of this discovery; coupled Ising models are *not* in the Ising universality class, in contrast to what a too extended application of universality would suggest.

Note also how much exact solvability is a delicate property. When written in terms of Ising spins, the Eight Vertex or the Ashkin Teller models look almost identical; however, an exact solution is known only for the second one. One expects that the exponents for the model (2.7) are non constant function of λ also in the non solvable cases.

The understanding of the universality issue for such models grew out from a number of authors in the Seventies and early Eighties, see *e.g.* [31],[33],[38],[26] and several others; it was proposed that the exponents, though model-dependent, verify a set of universal *extended scaling relations* allowing one to express *every exponents of a single model in terms of any one of them*. Some example of such relations are

$$X_- = \frac{1}{X_+} \qquad \nu = \frac{1}{2 - X_+} \qquad \alpha = 2 - 2\nu \tag{2.13}$$

where X_+ and X_- are the energy and crossover exponents for the correlations defined in (2.9), and ν, α are the exponents for the correlation length and the specific heat. Such exponents depend on the choice of V but the relations are model-independent and, once one exponent is fixed (say X_+) all the others are determined. The first of (2.13) was proposed by Kadanoff [31], the second by Kadanoff and Wegner [33] and the third is one of the hyperscaling relations [58]. Of course some of the relations can be checked in certain solvable models; it is the case of the relation $\alpha = 2 - 2\nu$ which was checked in the case of the Eight Vertex model (see Eq.(10.12.24) of [5]). It is worth however to remark that, even when there is an exact solution, not all the exponents can be computed; for instance the exponents $X_\pm$ for the Eight Vertex model cannot be computed.

The mathematical proof of such universal relations (as well as other ones which have been conjectured for such systems) has shown to be a rather challenging problem. Several attempts in the last thirty years have been devoted to their proof [48], [45],[61] ,[55], using a variety of methods ranging from operator product expansions, perturbation theory, Renormalization Group, bosonization and several others. It is common to all such approaches to start from a formal continuum limit in which extra symmetries are verified. However strictly speaking such a formal limit is plagued by contact divergences which were absent in the original lattice model. Moreover lattice effects destroy such symmetries and

change the exponents, and it is not clear at all why the relations between exponents should be true also when such symmetries are violated. Indeed, while the assumption of a continuum limit description of planar lattice model is very powerful, it is well known that a mathematical justification of it is very difficult, see *e.g.* [53].

Our main result is the proof of the extended scaling relations (2.13) (and several others) for coupled Ising models (2.7) and for related models as well.

The proof applies to solvable or non solvable cases and it is based on the new methods that have been introduced in [54] and [40] to study the critical properties of perturbations of the 2D Ising model. We will describe the main steps leading to it in the following sections.

3. Grassmann Integrals Representation

The starting point of the proof is the representation of the partition function of spin models with Hamiltonian (2.1) or (2.7) in terms of (non gaussian) Grassmann integrals; such representation, following from the works of Schultz, Lieb and Mattis [52], Hurst and Green [28] and Samuel [51], was known since a long time but only recently the progresses in Constructive Quantum field Theory [25] made it an useful starting point for the analysis of spin models.

The partition function of the Ising model with *periodic* boundary conditions can be written in terms of *Pfaffians*, see eq.(V.2.12) of [44]

$$Z = \sum_{\sigma} e^{-\beta H_J(\sigma)} = \sum_{\gamma} (-1)^{\delta_\gamma} Z_\gamma \tag{3.1}$$

where $\gamma = (\varepsilon, \varepsilon')$; $\varepsilon, \varepsilon' = \pm 1$; moreover $\delta_{+,+} = 1$ and $\delta_{-,+} = \delta_{+,-} = \delta_{-,-} = 2$ and

$$Z_\gamma = (-1)^{L^2} \frac{1}{2} (2 \cosh \beta J)^{L^2} \mathrm{Pf} A_\gamma \tag{3.2}$$

and A_γ matrices with elements $(A_i)_{\mathbf{x},j;\mathbf{y},k}$, with $\mathbf{x}, \mathbf{y} \in \Lambda$, $j, k = 1, \dots, 6$, given by:

$$(A_\gamma)_{\mathbf{x};\mathbf{x}} = \begin{pmatrix} 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 & -1 & 0 \end{pmatrix} \tag{3.3}$$

and, if $t = \tanh(\beta J)$, $\left((A_\gamma)_{\mathbf{x};\mathbf{x}+\mathbf{e_1}}\right)_{i,j} = t\delta_{i,1}\delta_{j,2}$, $\left((A_i)_{\mathbf{x};\mathbf{x}+\mathbf{e_0}}\right)_{i,j} = t\delta_{i,2}\delta_{j,1}$, $(A_i)_{\mathbf{x};\mathbf{x}+\mathbf{e_1}} = -({A_i}^T)_{\mathbf{x}+\mathbf{e_1};\mathbf{x}}$, $(A_i)_{\mathbf{x};\mathbf{x}+\mathbf{e_0}} = -({A_i}^T)_{\mathbf{x}+\mathbf{e_0};\mathbf{x}}$; moreover

$$\begin{aligned} (A_\gamma)_{(L,x_0);(1,x_0)} &= -(A_\gamma^T)_{(1,x_0);(L,x_0)} = \varepsilon (A_\gamma)_{(1,x_0);(2,x_0)} \\ (A_\gamma)_{(x,L);(x,1)} &= -(A_\gamma^T)_{(x,1);(x,L)} = \varepsilon' (A_\gamma)_{(x,1);(x,2)} \end{aligned} \tag{3.4}$$

and in all the other cases the matrices $(A_i)_{\mathbf{x},\mathbf{y}}$ are identically zero.

The Ising model partition function (3.1) can be conveniently rewritten in terms of *Grassmann integrals*. We recall that (see [17] for a detailed introduction), given a set of Grassmann variables η_α, with α belonging to some finite set A and $\{\eta_\alpha, \eta_{\alpha'}\} = \eta_\alpha\eta_{\alpha'} + \eta_{\alpha'}\eta_\alpha = 0$, a *Grassmann integral* is a linear functional $d\eta_\alpha$ such that

$$\int d\eta_\alpha = 0, \qquad \int d\eta_\alpha \eta_\alpha = 1 \tag{3.5}$$

and the integral of any analytic function can be obtained by linearity. A Pfaffian can be conveniently written in terms of Grassmann variables; indeed, given a $(2n) \times (2n)$ antisymmetric matrix A

$$\mathrm{Pf} A = (-1)^n \int d\eta_1 \cdots d\eta_{2n} e^{\frac{1}{2}\sum_{i,j}\eta_i A_{ij}\eta_j}$$

By using the above representation it is therefore straightforward to write (3.2) as:

$$Z = \frac{1}{2}(2\cosh\beta J)^{L^2}\sum_\gamma (-1)^{\delta_\gamma}\int \prod_{\mathbf{x}\in\Lambda} d\overline{H}_{\mathbf{x}} dH_{\mathbf{x}} d\overline{V}_{\mathbf{x}} dV_{\mathbf{x}}^\gamma d\overline{T}_{\mathbf{x}} dT_{\mathbf{x}} e^{\bar S^\gamma} \tag{3.6}$$

where $\overline{H}_{\mathbf{x}}, H_{\mathbf{x}}, \overline{V}_{\mathbf{x}}, V_{\mathbf{x}}, \overline{T}_{\mathbf{x}}, T_{\mathbf{x}}$ is a *finite set* of Grassmannian variables with ε–periodic resp. ε'–periodic boundary conditions in vertical resp. horizontal direction and

$$\begin{aligned}\bar S^\gamma &= t\sum_{\mathbf{x}}\Big[\overline{H}_{\mathbf{x}} H_{\mathbf{x}+\mathbf{e}_1} + \overline{V}_{\mathbf{x}} V_{\mathbf{x}+\mathbf{e}_0}\Big] + \\ &+ \sum_{\mathbf{x}}\Big[\overline{V}_{\mathbf{x}}\overline{H}_{\mathbf{x}} + \overline{H}_{\mathbf{x}} T_{\mathbf{x}} + V_{\mathbf{x}} H_{\mathbf{x}} + H_{\mathbf{x}}\overline{T}_{\mathbf{x}} + T_{\mathbf{x}}\overline{V}_{\mathbf{x}} + \overline{T}_{\mathbf{x}} V_{\mathbf{x}} + \overline{T}_{\mathbf{x}} T_{\mathbf{x}}\Big]\end{aligned} \tag{3.7}$$

The $T, \bar T$ variables, which appear only in the diagonal elements, can be easily integrated out; by a suitable changes of variables [29] and partial integrations, the integrals for Z_γ can be more conveniently expressed as (if $\gamma = (-,-)$ for definiteness)

$$(Z_\gamma)^2 = \mathcal{N}_1 \int \prod_{\omega=\pm,\mathbf{k}} d\psi^+_{\mathbf{k},\omega} d\psi^-_{\mathbf{k},\omega} e^{-\frac{Z}{L^2}\sum_{\mathbf{k}}\psi^+_{\mathbf{k},\omega} A_{\mathbf{k}} \psi^-_{\mathbf{k},\omega}} = \mathcal{N}_2 \int P_{Z,\mu}(d\psi) \tag{3.8}$$

where $\mathcal{N}_1, \mathcal{N}_2$ are constants, $\psi^\pm_{\mathbf{k},\omega}$, $\mathbf{k} \in \mathcal{D}$ and $\omega = \pm 1$ are a finite set of Grassmann variables, $\mathcal{D}$ is the set of $\mathbf{k} = (k_0, k_1)$ such that $k_0 = \frac{2\pi}{L}(n_0 + \frac{1}{2})$ and $k_1 = \frac{2\pi}{L}(n_0 + \frac{1}{2})$ for $n_0, n_1 = -\frac{L}{2}, ..., \frac{L}{2} - 1$, L an even integer and

$$A_{\mathbf{k}} = \begin{pmatrix} -i\sin k_0 + \sin k + \mu_{11}(\mathbf{k}) & -\mu + \mu_{12}(\mathbf{k}) \\ -\mu + \mu_{21}(\mathbf{k}) & -i\sin k_0 - \sin k_1 + \mu_{22}(\mathbf{k}) \end{pmatrix} \tag{3.9}$$

with $\mu = O(|\beta - \beta_c|)$, $\tanh \beta_c J = \sqrt{2} - 1$, $Z = O(1)$, $\mu_{ij} = O(\mathbf{k}^2)$.

Note that $P_{Z_1,\mu_1}(d\psi)$ is a Grassmann *Gaussian* integration; the exact solvability of the Ising model is reflected from the fact that the partition function is expressed in terms of a gaussian integral.

Let us consider now the coupled Ising model (2.7); we will be interested in particular in the specific heat C_v and the energy $\epsilon = +$ and cross-over ($\epsilon = -$) correlations, defined as in (2.9). Starting from (3.7) such correlations can be written as sums of functional derivatives (with respect to A^ϵ, $\epsilon = +$ for the energy and $\epsilon = -$ for the crossover) of Grassmann integrals with different boundary conditions; in the thermodynamic limit and $\beta \neq \beta_c$ it is sufficient to consider only one of them which is given by, in the case $J = J'$ (for definiteness)

$$Z(A) = \int P_{Z_1,\mu_1}(d\psi) e^{L^2\mathcal{N} + \mathcal{V}^{(1)}(\sqrt{Z_1}\psi) + \mathcal{B}^{(1)}(\sqrt{Z_1}\psi, A)} , \tag{3.10}$$

where $\mathcal{N}$ is a constant, $\psi^{\pm}_{\mathbf{x},\omega}$ is a finite set of Grassmann variables, $P_{Z_1,\mu_1}(d\psi)$ is given by (3.8) with $Z = Z_1$ and with $\mu_1 = O(|t - t_c|)$, $t = \tanh \beta J$, $t_c = \tanh \beta_c J = \sqrt{2} - 1 - \zeta$ and

$$\mathcal{V}^{(1)}(\psi) = \zeta_1 \sum_{\mathbf{x},\omega=\pm} \psi^+_{\mathbf{x},\omega}\psi^-_{\mathbf{x},-\omega} + \lambda_1 \sum_{\mathbf{x}} \psi^+_{\mathbf{x},+}\psi^-_{\mathbf{x},+}\psi^+_{\mathbf{x},-}\psi^-_{\mathbf{x},-} + R_1(\psi) \tag{3.11}$$

$$\mathcal{B}^{(1)}(\psi, A) = \sum A_{\epsilon,\mathbf{x}} O^\epsilon_{\mathbf{x}} + R_2(A, \psi) \tag{3.12}$$

with $\zeta_1 = O(\zeta), \lambda_1 = O(\lambda)$; R_1 is a sum of monomials in ψ more than quartic in ψ or quartic with at least a derivative and R_2 is a sum of monomials in A, ψ more than quadratic in ψ or quadratic with at least a derivative; finally

$$O^+_{\mathbf{x}} = \psi^+_{\mathbf{x},+}\psi^-_{\mathbf{x},-} + \psi^+_{\mathbf{x},-}\psi^-_{\mathbf{x},+} \qquad O^-_{\mathbf{x}} = i[\psi^+_{\mathbf{x},+}\psi^+_{\mathbf{x},-} + \psi^-_{\mathbf{x},+}\psi^-_{\mathbf{x},-}] \tag{3.13}$$

The parameter ζ (usually called a *counterterm*) has to be chosen so that β_c is the critical temperature (in general the critical temperature in (2.7) is different with respect to the Ising one).

While the Grassmann integral (3.8) appearing in the computation of the partition function of the Ising model is Gaussian, the Grassmann integral (3.10) for (2.1) is *non Gaussian* and it cannot be explicitly performed. In order to evaluate it, the simplest possibility is to expand the exponential integrated in (3.10) in Taylor series and use the Wick rule

$$\int P_{Z_1,\mu_1}(d\psi)\psi^-_{\mathbf{x}_1,\omega_1}...\psi^-_{\mathbf{x}_n,\omega_n}\psi^+_{\mathbf{y}_1,\omega'_1}...\psi^+_{\mathbf{y}_n,\omega'_n} = \sum_\pi (-1)^{p_\pi} \prod_{i=1}^n g_{\omega_1,\omega_{\pi(i)}}(\mathbf{x}_i - \mathbf{y}_{\pi(i)}) \tag{3.14}$$

where the sum is over all the permutations $\pi = (\pi(1), .., \pi(n))$ of $(1, ..., n)$, p_π is the parity with respect to the fundamental permutation and $g_{\omega,\omega'}(\mathbf{x} - \mathbf{y})$ is the *propagator*

$$g_{\omega,\omega'}(\mathbf{x} - \mathbf{y}) = \frac{1}{Z_1 L^2} \sum_{\mathbf{k}\in\mathcal{D}} e^{-i\mathbf{k}(\mathbf{x}-\mathbf{y})} [A^{-1}_{\mathbf{k}}]_{\omega,\omega'} \tag{3.15}$$

In this way Grassmann integrals representing the correlations of (2.7) are written as power series (in λ and ζ); they are *analytic* uniformly in L as function of λ, ζ with a radius of convergence *shrinking to zero* as $\beta \to \beta_c$. This means that such series *cannot be of any help* for the understanding of the critical behavior of the spin model (2.7). Indeed in the power series expansion the n-th contribution is given by a sum of coefficients, which can be conveniently graphically expressed in terms of *Feynman diagrams*. Each term is given by the integral of a product of propagators which, by (3.9) and (3.15), at the critical temperature $\beta = \beta_c$ have slow decay properties at large distances; therefore there are coefficients in the power series expansion which are unbounded as $\beta \to \beta_c$ and $L \to \infty$. This slow decay of the propagator is due to the singularity of $A_{\mathbf{k}}^{-1}$ at $\mathbf{k} = (0, 0)$ when $\mu = 0$.

We recall now that, in the absence of interaction, Dirac fermions are described by the *Dirac equation*, see *e.g.* [29], introduced in 1928 by Dirac to describe elementary spin 1/2 particles (like electrons) at high energies, in agreement with the principles of quantum mechanics and special relativity. It turns out that. in presence of interactions and with a lattice regularization *interacting Dirac fermions in $d = 1 + 1$ are described by Grassmann integrals somewhat similar to (3.10).*

There is therefore a remarkable connection between two apparently completely unrelated objects, namely relativistic quantum particles appearing in high energy physics, and Ising models, which are a classical description for a magnet. This is a further example of how apparently completely unrelated phenomena can reveal astonishing similarities at a deep mathematical level. We will use such a connection to apply, to the problem of the critical behavior of spin model (2.7), the powerful methods developed in the context of Constructive Quantum Field Theory [25]; by them a *resummation* of the series expansion can be found from which information on the critical behavior can be extracted.

4. Renormalization Group and Multiscale Decomposition

Starting from the Grassmann integral representation (3.10) the following Theorem has been proved

Theorem 4.1. *(Mastropietro [39],[40]) The coupled Ising model (2.7) with $J = J'$ and λ small enough is critical at $\tanh \beta_c J = \sqrt{2} - 1 + O(\lambda)$ and the specific heat exists and is such that*

$$C_v \sim -\frac{1}{\alpha}[1 - |\beta - \beta_c|^{-\alpha}] \tag{4.1}$$

with $\alpha = O(\lambda)$. If $\beta \neq \beta_c$ the energy and crossover correlations $G^{\epsilon}_{\beta}(\mathbf{x}-\mathbf{y})$, $\epsilon = \pm$ (2.9),(2.10) decay faster than any power of $\xi^{-1}|\mathbf{x}-\mathbf{y}|$, with

$$\xi^{-1} \sim |\beta - \beta_c|^{\nu} \tag{4.2}$$

with $\nu = 1 + O(\lambda)$. Finally

$$G^{\epsilon}_{\beta_c}(\mathbf{x}-\mathbf{y}) \sim \frac{1}{|\mathbf{x}-\mathbf{y}|^{2X_\epsilon}} \text{ , as } |\mathbf{x}-\mathbf{y}| \to \infty \text{ ,} \tag{4.3}$$

with $X_{\pm} = 1 + O(\lambda)$.

The quartic interaction in (2.7) has two main effects. The first one is simply to change the value of the critical temperature. The second and more dramatic is that the critical behavior is modified even by an arbitrarily small interaction. Indeed the exponents $\alpha, X_{\pm}, \nu$ are expressed in terms of convergent expansions whose lowest order coefficients can be explicitly computed. It is found, in the case of the Askhin-Teller (2.11) or Eight Vertex model (2.12)

$$\alpha = 2a_1\lambda + O(\lambda^2), \quad X_{\pm} = 1 \mp a_1\lambda + O(\lambda^2), \quad \nu = 1 - a_1\lambda + O(\lambda^2) \tag{4.4}$$

with a_1 a suitable positive constant. Therefore the logarithmic singularity in the specific heat of the Ising model is changed by the interaction into a power law singularity if $\lambda > 0$; if $\lambda < 0$ the specific heat is indeed continuous, but higher order derivatives of the free energy are singular. In order to establish critical behavior the specific heat is evaluated at $\beta \neq \beta_c$ and $L = \infty$ and we verify that it (or some of its derivatives) has a singular behavior as $\beta \to \beta_c$; the limit $L \to \infty$ is not taken directly at $\beta = \beta_c$.

Theorem (4.1) proves for the first time that the critical exponents in generic *non solvable* coupled Ising models (2.7) are non trivial functions of the coupling. From (4.4), the scaling relations (2.13) are verified if the expansion is truncated at first order; however the complexity of the expansions makes essentially impossible to prove the universal relations directly from the series by an analysis at all orders.

Let us consider now what happens when $J \neq J'$, in the case of the anisotropic Ashkin-Teller model for definiteness. If $J - J'$ is large (strong anisotropy) the two Ising subsystems have very different critical temperatures, hence one can expect that if one system is almost critical the second one will be out of criticality: the system is expected then to be in the Ising universality class. On the other hand if $J = J'$ the exponents are non trivial functions of λ, as shown in the previous theorem, and the system is not in the Ising class; how the crossover is realized for $J - J'$ small is clarified by the following theorem.

Theorem 4.2. *(Giuliani,Mastropietro [23],[24]) In the case of the anisotropic Ashkin-Teller model (2.7),(2.11) ($J \neq J'$) there are two critical temperatures, β_c^+ and β_c^- such that*

$$|\beta_c^- - \beta_c^+| \sim |J - J'|^{X_T} \tag{4.5}$$

with $X_T = 1 + O(\lambda)$ *and*

$$C_v \sim -\Delta^\alpha \log \frac{|\beta - \beta_c^-| \cdot |\beta - \beta_c^+|}{\Delta^2} \tag{4.6}$$

where $2\Delta^2 = (\beta - \beta_c^-)^2 + (\beta - \beta_c^+)^2$ *and* $\alpha = O(\lambda)$.

By the above theorem we can see that the anisotropic Ashkin-Teller model is in the class of universality of the Ising model for any nonvanishing value of $J - J'$: the specific heat has the same logarithmic singularity as in the Ising model, and the $X_\pm$ and ν exponents are the Ising ones. However critical exponents which are non trivial function of the coupling λ appear even if we are in the Ising universality class: the difference between the two critical temperatures rescales with an anomalous exponent in the isotropic limit $|\beta_c^+ - \beta_c^-| \sim |J - J'|^{X_T}$. Therefore, the ratio of the difference of the critical temperatures when $\lambda \neq 0$ or $\lambda = 0$ is vanishing or diverging as $J \to J'$ depending on the sign of λ, with a power law driven by a *transition exponent* X_T, whose existence was overlooked in the physical literature.

Let us give the main ideas of the proof of Theorem 4.1, which is based on an approach known as *Renormalization Group*, which produces a *resummation* of the power series expansion of (3.10) which is well defined for β close to β_c uniformly L. Developed by Wilson [59] for Statistical Physics or Quantum Field Theory, the *ideas* of Renormalization Group are used also in many mathematical studies and often go under the name "multiscale analysis"; for instance, such ideas are used in the proof of the pointwise convergence of Fourier series on the circle, or in the convergence of Lindstedt series for KAM tori, see [19].

We follow the application of Wilsonian ideas to functional integrals like (3.10) due to Polchinski [47] and Gallavotti [20]. The starting point (see *e.g.* [42] for a tutorial introduction) is the *addition property* for Gaussian Grassmann integrals; if $P(d\psi), P(d\psi^1), P(d\psi^2)$ are Gaussian Grassmann integrations with propagator $g(\mathbf{k}), g^{(1)}(\mathbf{k}), g^{(2)}(\mathbf{k})$, with $g^{(1)}(\mathbf{k}) + g^{(1)}(\mathbf{k}) = g(\mathbf{k})$ then the integration can be equivalently rewritten as

$$\int P(d\psi) F(\psi) = \int P(d\psi^{(2)}) \int P(d\psi^{(1)}) F(\psi^{(1)} + \psi^{(2)}) \tag{4.7}$$

Let T^1 be the one dimensional torus, $||k - k'||_{T^1}$ the usual distance between k and k' in T^1 and $||k|| = ||k - 0||$. We introduce a decomposition of the unity

$$1 = f_1(\mathbf{k}) + \sum_{h=-\infty}^{0} f_h(\mathbf{k}) \tag{4.8}$$

with $f_h(\mathbf{k})$: if $h \leq 0$ a smooth compact support function with support $\{\frac{\pi}{4} 2^{h-1} \leq |\mathbf{k}| \leq \frac{\pi}{4} 2^{h+1}$; if $h = 1$ $f_1(\mathbf{k}) = 0$ for $|\mathbf{k}| \leq \frac{\pi}{4} 2^{-1}$ and $f_1(\mathbf{k}) = 1$ for $|\mathbf{k}| \geq \frac{\pi}{4} 2$. We define also

$$\chi_h(\mathbf{k}) = \sum_{k=-\infty}^{h} f_h(\mathbf{k}) \tag{4.9}$$

which is vanishing for $|\mathbf{k}| \geq \frac{\pi}{4} 2^{h+1}$. Therefore, by applying (4.7), we can write $Z(A)$ (3.10) as

$$\int P_{Z_1,\mu_1}(d\psi^{\leq 0}) \int P_{Z_1,\mu_1}(d\psi^{(1)}) e^{L^2\mathcal{N}+\mathcal{V}^{(1)}(\sqrt{Z_1}\psi)+\mathcal{B}^{(1)}(\sqrt{Z_1}(\psi^{(\leq 0)}+\psi^{(1)},A)}$$
$$= e^{S^{(0)}(A)} \int P_{Z_1,\mu_1}(d\psi^{\leq 0}) e^{L^2\mathcal{N}^{(0)}+\mathcal{V}^{(0)}(\sqrt{Z_1}\psi^{(\leq 0)})+\mathcal{B}^{(0)}(\sqrt{Z_1}\psi^{(\leq 0)},A)}, \quad (4.10)$$

where $\mathcal{V}^{(0)}, \mathcal{B}^{(0)}, S^{(0)}$ is sum over all monomials multiplied by suitable kernels. The advantage of this is that, while the propagator of $P_{Z_1,\mu_1}(d\psi)$ has slow decay properties, the propagator $P_{Z_1,\mu_1}(d\psi^{(1)})$ decays faster than any power, and the result of the integration is well defined (the kernels are analytic for small λ, ζ uniformly in L, β).

We can iterate such a procedure and, after some rescaling and other manipulations, one defines a sequence of *effective potentials* $\mathcal{V}^{(h)}$, *effective sources* $\mathcal{B}^{(h)}, S^{(h)}$ and a sequence of constants $Z_h, \mu_h, \mathcal{N}_h$, $h = 0, -1, , ..$ such that

$$Z(A) = e^{S^{(h)}(A)} \int P_{Z_{h-1},\mu_{h-1}}(d\psi^{(\leq h)}) e^{L^2\mathcal{N}_h+\mathcal{V}^{(h)}(\sqrt{Z_{h-1}}\psi^{(\leq h)})+\mathcal{B}^{(h)}(\sqrt{Z_{h-1}}\psi^{(\leq h)},A)}, \quad (4.11)$$

where $\psi^{(\leq h)} = \sum_{j=-\infty}^{h} \psi^{(j)}$ and $P_{Z_h,\mu_h}(d\psi^{(\leq h)})$ is the *Gaussian Grassmann integration* with propagator $g^{(\leq h)}(\mathbf{x})$, with Fourier transform given by

$$\widehat{g}^{(\leq h)}(\mathbf{k}) = \frac{\chi_h(\mathbf{k})}{Z_h} \begin{pmatrix} -i\sin k_0 + \sin k + \mu_{++} & -\mu_h - \mu_{-+} \\ -\mu_h - \mu_{+-} & -i\sin k_0 - \sin k_1 + \mu_{--} \end{pmatrix}^{-1} \quad (4.12)$$

The *effective interaction* $\mathcal{V}^{(h)}(\psi)$ is a sum over monomials in the Grassmann variables

$$\mathcal{V}^{(h)}(\psi^{(\leq h)}) = \gamma^h \zeta_h F_\nu^{(h)} + \lambda_h F_\lambda^{(h)} + R_h(\psi^{(\leq h)}), \quad (4.13)$$

where

$$F_\nu^{(h)} = \frac{1}{L^2} \sum_{\omega=\pm} \sum_{\mathbf{k}} \psi^{(\leq h)+}_{\mathbf{k},\omega} \psi^{(\leq h)-}_{\mathbf{k},-\omega}, \quad (4.14)$$

$$F_\lambda^{(\leq h)} = \frac{1}{L^8} \sum_{\mathbf{k}_1,\ldots,\mathbf{k}_4} \psi^{(\leq h)+}_{\mathbf{k}_1,+} \psi^{(\leq h)+}_{\mathbf{k}_3,-} \psi^{(\leq h)-}_{\mathbf{k}_2,+} \psi^{(\leq h)-}_{\mathbf{k}_4,-} \delta(\mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3 - \mathbf{k}_4).$$

and R^h contains sum of monomials with more than four fields, or quartic with at least a derivative, or bilinear with at least two derivatives. In the same way

$$\mathcal{B}^{(h)}(\sqrt{Z_{h-1}}\psi^{(\leq h)}, A) = \sum_{\epsilon=\pm,\mathbf{x}} Z_{h-1}^{(\epsilon)} A_{\mathbf{x}}^{\epsilon} O_{\mathbf{x}}^{(\leq h)\epsilon} + \bar{R}_h, \quad (4.15)$$

where $O^{\pm}$ is given by (3.13) and $\bar{R}_h$ contains terms more than quadratic, or quadratic with a derivative.

The above procedure has the effect that, after the integration of the fields $\psi^{(1)}, ..., \psi^{(h+1)}$, $Z(A)$ is expressed by a functional integral similar to (3.10), with the difference that the fields have support in a smaller momentum region and have, in general, *renormalized* masses, velocities and wave function renormalization; in addition, the interaction is replaced by an effective interaction $\mathcal{V}^{(h)}$ which is typically sum of monomials of any degree in the fields $\psi^{(\le h)}$ and A. To each monomial in the effective potential is associated a *scaling dimension*. If only a finite set of monomials have non negative scaling dimension the theory is said *renormalizable*; this is what happens in the present case as the *scaling dimension* is given, for the monomials with n ψ-fields and m A-fields, by

$$D = 2 - \frac{n}{2} - m \; . \tag{4.16}$$

so that only the monomials with $(n, m) = (2, 0); (2, 1); (4, 0); (0, 2)$ have non positive dimension.

The crucial point is that the kernels of the effective potentials $\mathcal{V}^{(h)}, \mathcal{B}^{(h)}, S^{(h)}$ can be written as a power series expansions in the effective couplings $\{\lambda_k, \zeta_k\}_{k\ge h}$ which is *convergent uniformly in L and $\beta - \beta_c$*, provided that $\{\lambda_k, \zeta_k\}_{k\ge h}$ are small enough. The proof of this remarkable property is based on the *Gallavotti-Nicolo' tree* expansion [21], the *Battle-Bridges-Federbush formula* [3] for the truncated fermionic expectation together with the *Gram bounds* for determinants [15]; it is technically similar to the analysis performed in constructive Quantum Field Theory by Gawedsky and Kupianen [22] and Feldman, Magnen, Rivasseau and Seneor [16] for the Gross-Neveu model, or by Lesniewski [35] for the Yukawa model [35]. Note that the multiscale integration procedure has replaced an expansion in λ, ζ which, as discussed at the end of the previous section, has *unbounded* coefficients, with an expansion in $\{\lambda_k, \zeta_k\}_{k\ge h}$ which has finite radius of convergence uniformly in L and $\beta - \beta_c$.

There are however two main differences between our functional integral (3.10) with respect to fermionic $d = 1 + 1$ Quantum Field Theories like the Gross-Neveu or the Yukawa models. The first is that, as it is usually said, (3.10) poses an *infrared* (that is, related to the divergence of the propagator at low momenta when $\beta = \beta_c$) and not an *ultraviolet* problem (related to the slow decay of the propagator for large momenta).

The second and more crucial one is that the theory (3.10) is renormalizable but *not* asymptotically free, as it is the case for the Gross-Neveu model. In the case we are discussing here, asymptotic freedom would mean that $\lambda_h, \zeta_h \to 0$ as $h \to -\infty$; this would ensure that, by choosing λ small enough, the expansion in terms of the effective coupling would be convergent. However this is *not* what happens in the case (3.10), in which $\lambda_h \to \lambda_{-\infty}$, with $\lambda_{-\infty}(\lambda) = \lambda_1 + O(\lambda^2)$ an analytic non trivial function of λ. In the Renormalization Group language, one says that there is *a line of fixed points*. Such kinds of models are much harder to be constructed with respect to the asymptotically free ones: one has to exploit non trivial cancellations in the expansions at *all orders* in the renormalized expansion; this is a crucial difference with respect to the asymptotically free

models in which a second order computation is enough for establishing the nature of the flow of the effective coupling.

The first example of rigorous construction of a model of this kind was in [6, 7] and it regards the Jellium model in 1D, describing interacting non relativistic fermions in the continuum. The cancellations were proved using an indirect argument based on comparison with the *exact solution* of the Luttinger model found by Mattis and Lieb [43]. Later on, other models with lines of fixed points were constructed *without any use of exact solutions*, using a technique, developed in [14], capable of combining Ward Identities based on local symmetries with Renormalization Group methods; the main problem to face is that the momentum cut-off breaks local symmetries producing additional terms in the Ward Identities which can be however rigorously taken into account.

An important observation which will play an important role in the proof of the universal relations (see below) is that the propagator of the field $\psi^{(h)}$ can be written, for $h \leq 0$, as

$$g^{(h)}(\mathbf{x}-\mathbf{y}) = g_T^{(h)}(\mathbf{x}-\mathbf{y}) + r^{(\leq h)}(\mathbf{x}-\mathbf{y}) \; , \tag{4.17}$$

where

$$g_T^{(h)}(\mathbf{x}-\mathbf{y}) = \frac{1}{L^2}\sum_{\mathbf{k}} e^{-i\mathbf{k}(\mathbf{x}-\mathbf{y})}\frac{1}{Z_h}T_h^{-1}(\mathbf{k}) \; , \tag{4.18}$$

$$T_h(\mathbf{k}) = f_h(\mathbf{k})\begin{pmatrix} -ik_0+k & -\mu_h \\ \mu_h & -ik_0-k \end{pmatrix} \tag{4.19}$$

and $r^{(h)}(\mathbf{x}-\mathbf{y})$ verifying for large distances the same bound as $g^{(h)}(\mathbf{x},\mathbf{y})$ times an extra 2^h. The above decomposition means that the single scale propagator is identical to a "relativistic" one (see the following section) up corrections which are smaller and smaller as $h \to -\infty$.

In order to prove that $\{\lambda_k, \zeta_k\}_{k\geq h}$ remain inside the radius of convergence one considers a recursive equation (whose r.h.s. is called *Beta function*)

$$\lambda_{j-1} = \lambda_j + \beta_\lambda^{(j)}(\lambda_j, ..., \lambda_0) + \bar\beta_\lambda^{(j)}(\lambda_j, \zeta_j; ...; \lambda_0, \zeta_0) + O(\bar\lambda_j 2^{\vartheta j}) \; , \tag{4.20}$$

where $0 < \vartheta < 1$ is a constant, $\bar\lambda_j = \max_{k\geq j}|\lambda_k|$, $\beta_\lambda^{(j)}$, $\bar\beta_\lambda^{(j)}$ are μ_1-independent and expressed by a *convergent* expansion in $\lambda_j, \zeta_j .., \lambda_0, \zeta_0$; moreover by definition $\beta_\lambda^{(j)}(\lambda_j, ..., \lambda_0)$ is sum of terms in which only the propagators $g_T^{(h)}$ (4.18) appear (the terms containing $r^{(j)}$ are included in the last term in the r.h.s. of(4.20)) and $\bar\beta_\lambda^{(j)}$ vanishes if at least one of the ζ_k is zero. Remarkable cancellations in the beta functions, expressed by the following bound

$$|\beta_\lambda^{(j)}(\lambda_j, ..., \lambda_j)| \leq C|\lambda_j|^2 2^{\theta j} \tag{4.21}$$

for suitable positive constants C and $\theta < 1$, and the fact that, for a suitable choice of $\zeta_1 = O(\lambda)$, $\zeta_j = O(2^{\theta j}\lambda)$ and therefore $\bar\beta_\lambda^{(j)} = O(2^{\theta j}\lambda^2)$, imply that

$$\lambda_j \to \lambda_{-\infty}(\lambda) = \lambda_1 + O(\lambda^2) \tag{4.22}$$

The critical exponents are found by the beta function for the effective renormalizations; we can write

$$\frac{Z_{j-1}}{Z_j} = 1 + \beta_z^{(j)}(\lambda_j, ..., \lambda_0) + \bar\beta_z^{(j)}(\lambda_j, \zeta_j; .., \lambda_0, \zeta_0) + O(\lambda 2^{\vartheta j}) \ , \tag{4.23}$$

with $\bar\beta_z^{(j)}$ vanishing if at least one of the ζ_k is zero so that $\bar\beta_z^{(j)} = O(\lambda 2^{\theta j})$. Finally

$$\beta_z^{(j)}(\lambda_j, ..., \lambda_0) = \beta_z^{(j)}(\lambda_{-\infty}, ..., \lambda_{-\infty}) + O(\lambda 2^{\theta h}) \ , \tag{4.24}$$

where $\beta_z^{(j)}(\lambda_{-\infty}, ..., \lambda_{-\infty})$ is by definition sum of terms in which only the propagators $g_T^{(h)}$ (4.18) appear (the terms containing $r^{(j)}$ are included in the second term in (4.24)). Similar equations hold for $Z_h^{(\pm)}, \mu_h$, with

$$\beta_\pm(\lambda_j, ..., \lambda_0) = \beta_\pm(\lambda_{-\infty}, ..., \lambda_{-\infty}) + O(\lambda 2^{\theta h}) \ . \tag{4.25}$$

so that, by defining

$$\eta_\pm = \log_2[1 + \beta_\pm^{(-\infty)}(\lambda_{-\infty}, ...\lambda_{-\infty})] \ , \tag{4.26}$$

and similar equations for the other exponent, we get for any $j \le 0$,

$$Z_j \sim 2^{\eta_z j} \quad \mu_j \sim \mu_1 2^{\eta_\mu j} \quad Z_j^{(\pm)} \sim 2^{\eta_\pm j} \tag{4.27}$$

The critical exponents in Theorem 4.1 are functions of $\lambda_{-\infty}$ only, as it is clear from (4.24), and are such that

$$X_\pm = 1 - \eta_\pm + \eta_z \qquad \eta_\mu = \eta_+ - \eta_z = 1 - X_+ \ . \tag{4.28}$$

If $\mu_1 \neq 0$ (that is, if the temperature is not the critical one), the correlations decay faster than any power with rate proportional to μ_{h^*}, where, if $[x]$ denotes the largest integer $\le x$, h^* is given by $h^* = \left[\frac{\log_2 |\mu_1|}{1+\eta_\mu}\right]$.

5. The Extended Scaling Relations

The exponents of the model (2.7) are written as convergent series so that they can be computed with arbitrary precision; at lowest orders, see (4.4), the relations (2.13) are verified, but to prove their validity at all orders directly from the expansions is essentially impossible due to the complexity of the series. Recently some of the universal relations have been proved.

Theorem 5.1. *(Benfatto,Falco,Mastropietro [10, 11]). Given the coupled Ising model with quartic interaction (1.4), with the same definitions as in Theorems 4.1 and 4.2 and λ small enough the following relations are true*

$$\begin{aligned} X_- &= \frac{1}{X_+} \qquad & \alpha &= \frac{2 - 2X_+}{2 - X_+} \ , \\ \nu &= \frac{1}{2 - X_+} \qquad & X_T &= \frac{2 - X_+}{2 - X_+^{-1}} \end{aligned} \tag{5.1}$$

The first three of the above relations were previously conjectured (see remarks after (2.13)) while the last one is completely new.

The idea of the proof of the above theorem is based on the introduction of a fermionic theory, defined on the *continuum* space and not on the lattice, whose correlations are the functional derivatives of the following Grassmann integral

$$\int P(d\psi^{(\le N)}) e^{V^{(N)}(\psi^{(\le N)}) + \sum_{\omega=\pm}\int d\mathbf{x}[\psi^+_{\mathbf{x},\omega}\phi^-_{\mathbf{x},\omega} + \psi^-_{\mathbf{x},\omega}\phi^+_{\mathbf{x},\omega}] + \int d\mathbf{x} j_\mu(\mathbf{x}) J_\mu(\mathbf{x})} \tag{5.2}$$

with $\mathbf{x} \in \Lambda$, $\Lambda \subset \mathbb{R}^2$, $\Lambda = [-L/2, L/2] \times [-L/2, L/2]$, and $P(d\psi^{(\le N)})$ is the fermionic gaussian integration with propagator

$$\widehat{g}^{(\le N)}(\mathbf{k}) = \frac{\chi_N(\mathbf{k})}{\not{\mathbf{k}}} \tag{5.3}$$

$\not{\mathbf{k}} = \gamma_0 k_0 + \gamma_1 k_1$, $\phi^\pm, J$ are *external fields* and

$$V^{(N)}(\psi^{(\le N)}) = \lambda_\infty \sum_{\mu=0,1} \int d\mathbf{x} d\mathbf{y} v(\mathbf{x}-\mathbf{y}) j_\mu(\mathbf{x}) j_\mu(\mathbf{y}) \tag{5.4}$$

with $j_\mu(\mathbf{x}) = \bar\psi_{\mathbf{x}} \gamma_\mu \psi_{\mathbf{x}}$, $\bar\psi = \psi^+\gamma_0$, $\psi^+ = (\psi^+_+, \psi^+_-)$, $\psi = (\psi^-_-, \psi^-_-)$ and $v(\mathbf{x}-\mathbf{y})$ a short range symmetric interaction with $\widehat{v}(0) = 1$ and $|\widehat{v}(\mathbf{p})| \le e^{-\kappa_0|\mathbf{p}|}$; moreover

$$\gamma_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \gamma_1 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \gamma_5 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{5.5}$$

The functional integral (5.2) is a *regularization* of the formal continuum limit of (3.10) in which the lattice step is sent to zero and $\beta = \beta_c$; note indeed the single-scale propagator of this model is coinciding with $g^{(h)}_T(\mathbf{x})$ in (4.17) when $\mu_h = 0$ and $\mathbf{k} \in \mathbb{R}^2$ (when $L = \infty$). Grassmann integrals similar to (5.2), with $v(\mathbf{x})$ replaced by a local interaction, appears in the construction of a Quantum Field Theory known as *Thirring model*, describing Dirac fermions with a local current-current interaction, see *e.g.* [9].

By a multiscale integration in the ultraviolet region one can perform safely the limit $N \to \infty$, see [41]. The multiscale integration for the infrared scales can be done exactly as described in the previous section for the model (3.10); the single scale propagator is given by the dominant propagator in (4.17) and, after the integration of $\psi^{(N)}, ..., \psi^{(j+1)}$, one obtains an equation very similar to (4.11), with the difference that, by the oddness of the free propagator, $\zeta_j = 0$ and

$$\lambda_{j-1} = \lambda_j + \beta^{(j)}_\lambda(\lambda_j, ...\lambda_0) + O(\bar\lambda_j^2 2^{\vartheta j}) \, , \tag{5.6}$$

with $\beta^{(j)}_\lambda(\lambda_j, ...\lambda_0)$ being *the same function appearing in (4.21)* for the model (2.7). Therefore we can prove that $\lambda_{-\infty} = \lambda_0 + O(\lambda_0^2)$; since $\lambda_0 = \lambda_\infty + O(\lambda_\infty^2)$, we have

$$\lambda_{-\infty} = h(\lambda_\infty) = \lambda_\infty + O(\lambda_\infty^2) \, , \tag{5.7}$$

for some analytic function $h(\lambda_\infty)$, invertible for λ_∞ small enough. Moreover

$$\frac{Z^{\pm}_{j-1}}{Z^{\pm}_{j}} = 1 + \beta^{(j)}_{\pm}(\lambda_j, ...\lambda_0) + O(\lambda_\infty 2^{\vartheta j}) \, , \tag{5.8}$$

with $\beta^{(j)}_{\pm}$ *exactly coinciding* with the functions appearing in (4.25); moreover $\eta_\pm = \log_2[1 + \beta^{(-\infty)}_{\pm}(\lambda_{-\infty}, ...\lambda_{-\infty})$.

While $\eta_\pm$ in the models (3.10) and (5.2) are the same as functions of $\lambda_{-\infty}$, of course they are completely different as function of the coupling λ_1 and λ_∞ appearing respectively in (3.10) and (5.2). However all the dependence on the model details is hidden in $\lambda_{-\infty}$: therefore if we call $\lambda'_j(\lambda)$ the effective couplings of the model (3.10) appearing in the previous section, the invertibility of $h(\lambda_\infty)$ implies that we can choose λ_∞ so that

$$\lambda_{-\infty} = h(\lambda_\infty) = \lambda'_{-\infty}(\lambda) \tag{5.9}$$

This implies that *the exponents in the models (2.7) and (5.2) are the same, provided that bare coupling λ_∞ is chosen properly.*

The point is now that the continuum fermionic theory (5.2) verifies extra exact identities with respect to the original spin Hamiltonian. Indeed we can perform the change of variables

$$\psi^{\pm}_{\mathbf{x}} \to e^{\pm i\alpha_{\mathbf{x}}}\psi^{\pm}_{\mathbf{x}} \tag{5.10}$$

in (5.2); the interaction (5.4) is invariant, while $P(d\psi)$ and the source terms are changed (and the Jacobian is 1). The derivative with respect to $\alpha_{\mathbf{x}}$ of (5.2) after the the change of variables $\psi^{\pm}_{\mathbf{x}} \to e^{\pm i\alpha_{\mathbf{x}}}\psi^{\pm}_{\mathbf{x}}$ is of course vanishing; therefore, by deriving with respect to $\alpha_{\mathbf{x}}$ and to the external fields we get, if $\langle ... \rangle$ are the correlations with respect to $P(d\psi^{(\le N)})e^{\mathcal{V}^{(N)}}$, the following *Ward Identity*

$$-i\sum_{\mu}\mathbf{p}_\mu\langle j_{\mu,\mathbf{p}}\psi_{\mathbf{k}}\bar\psi_{\mathbf{k}+\mathbf{p}}\rangle = \langle\psi_{\mathbf{k}}\bar\psi_{\mathbf{k}}\rangle - \langle\psi_{\mathbf{k}+\mathbf{p}}\bar\psi^{-}_{\mathbf{k}+\mathbf{p}}\rangle + \Delta_N(\mathbf{k},\mathbf{p}) \tag{5.11}$$

where

$$\Delta_N = \langle\delta j_{\mathbf{p}}\psi_{\mathbf{k}}\bar\psi_{\mathbf{k}+\mathbf{p}}\rangle \tag{5.12}$$

with

$$\delta_{\mathfrak{J}_{\mathbf{p}}} = \int d\mathbf{k}[(\chi_N^{-1}(\mathbf{k}+\mathbf{p})-1)(\not{k}+\not{p}) - (\chi_N^{-1}(\mathbf{k})-1)\ \not{k}]\bar\psi_{\mathbf{k}}\psi_{\mathbf{k}+\mathbf{p}} \tag{5.13}$$

An analogous expression is obtained for the axial current $\bar\psi\gamma_\mu\gamma_5\psi$.

If Ward Identities are derived from the ill-defined Grassmann integral (5.2) *without* momentum cut-off (that is, in the formal expression with $N = \infty$), one would get the same WI with $\Delta_N = 0$. On the contrary Δ_N is *not* vanishing in the limit $N \to \infty$ but, see [41]

$$\lim_{N\to\infty}\Delta_N(\mathbf{k},\mathbf{p}) = -i\tau\widehat{v}(\mathbf{p})\sum_{\mu}\mathbf{p}_\mu\langle j_{\mu,\mathbf{p}}\psi_{\mathbf{k},\omega}\bar\psi_{\mathbf{k}+\mathbf{p},\omega}\rangle \qquad \tau = \frac{\lambda_\infty}{4\pi} \tag{5.14}$$

An expression similar to (5.11) holds for the axial Ward Identity (the one obtained through the transformation $\psi^{\pm}_{\mathbf{x}} \to e^{\pm i\gamma_5\alpha_{\mathbf{x}}}\psi^{\pm}_{\mathbf{x}}$, with j_μ replaced by $j^5_\mu = \bar\psi\gamma_\mu\gamma_5\psi$ and τ replaced by $-\tau$; therefore, in the limit $N \to \infty$

$$-i\sum_\mu \mathbf{p}_\mu \langle j_{\mu,\mathbf{p}}\psi_{\mathbf{k},\omega}\bar\psi_{\mathbf{k}+\mathbf{p}}\rangle = A[\langle\psi_{\mathbf{k}}\bar\psi_{\mathbf{k}}\rangle - \langle\psi_{\mathbf{k}+\mathbf{p}}\bar\psi^-_{\mathbf{k}+\mathbf{p}}\rangle] \qquad (5.15)$$

$$-i\sum_\mu \mathbf{p}_\mu \langle j_{5,\mu,\mathbf{p}}\psi_{\mathbf{k},\omega}\bar\psi_{\mathbf{k}+\mathbf{p}}\rangle = \bar A[\langle\psi_{\mathbf{k}}\bar\psi_{\mathbf{k}}\rangle - \langle\psi_{\mathbf{k}+\mathbf{p}}\bar\psi^-_{\mathbf{k}+\mathbf{p}}\rangle]$$

with $A^{-1} = 1 - \tau\widehat{v}(\mathbf{p})$ and $A^{-1} = 1 + \tau\widehat{v}(\mathbf{p})$. From the Ward Identities one can write an equation for the correlations from which the exponents can be written in terms of τ and λ_∞

$$X_+ = 1 - \frac{1}{1+\tau}(\lambda_\infty/2\pi) \quad X_- = 1 + \frac{1}{1-\tau}(\lambda_\infty/2\pi) \qquad (5.16)$$

and from the above expression the relation (5.1) in Theorem (5.1) follows, as the exponents have simple expressions in λ_∞, as consequence of the linearity of τ. The other relations in Theorem 5.1 are proved in a similar way.

There are then two crucial points in the proof of the universal relations in Theorem 5.1. The first is that the exponents of the lattice theory (2.7) are equal to the ones of a continuum relativistic quantum field theory (5.2), *provided that* the coupling λ_∞ is chosen properly as a convergent series in λ with coefficients depending from all the details of the spin model (2.7). The second crucial point is that τ is linear in λ_∞, and this implies that the exponents have a simple expression as functions of λ_∞, from which the validity of the relations can be easily checked. Note that the validity of such a crucial property depends from the choice of a non local interaction in (5.4); with a somewhat more natural local interaction in (5.4) τ would be not linear in λ_∞, see [9].

The fact that $\Delta_N(\mathbf{k},\mathbf{p})$ is non vanishing removing the ultraviolet cut-off $N \to \infty$ is related to a *quantum anomaly*. In Quantum Field Theory anomalies are the breaking of classical symmetries by quantum mechanical radiative corrections; the classical Noether theorem is not verified in such cases due to the apparence of extra terms in the conservation laws. The linearity of τ in the bare coupling λ_∞ is the non-perturbative analogue, see [41], of a property called *anomaly non renormalization* in 4D Quantum Electrodynamic, proved at a perturbative level by Adler and Bardeen [1] with a careful analysis of the perturbative expansion.

6. Quantum Spin Chains

The Ising models seen previously describe magnets in which the dipoles are described by a spin variable $\sigma_{\mathbf{x}} = \pm$; in a more realistic description, given by the *Heisenberg model*, the spins, according to the rules of quantum mechanics,

are represented by operators. In one dimension the Heisenberg spin model has Hamiltonian

$$H = -\sum_{x=1}^{L-1}[J_1 S_x^1 S_{x+1}^1 + J_2 S_x^2 S_{x+1}^2 - hS_x^3] + \lambda \sum_{1\le x,y\le L} v(x-y) S_x^3 S_y^3 + U_L \quad (6.17)$$

where $S_x^\alpha = \sigma_x^\alpha/2$, $\alpha = 1,2,3$ and $[\sigma_x^\alpha, \sigma_y^\beta] = 0$ for $x \neq y$ while $[\sigma_x^\alpha, \sigma_y^\beta] = 2i\varepsilon_{\alpha\beta\gamma}\sigma_x^\gamma$; moreover $|v(x-y)| \le Ce^{-\kappa_0|x-y|}$ and the last term in (6.17) depend on the boundary conditions.

In the case of zero external magnetic field and nearest neighbor interaction, that is $v(x-y) = \delta_{|x-y|,1}/2$ and $h = 0$, the model is known as the *XYZ model* if $J_1 \neq J_2$ and it is exactly solvable; remarkably, it appears to be equivalent, with a suitable identification of the parameters, to the Eight Vertex model in the sense that, as was shown by Sutherland [57], the transfer matrix of the Eight Vertex model commutes with the Hamiltonian of the XYZ model.

Also in this case the model (6.17) can be exactly mapped onto a system of interacting fermions through the *Jordan-Wigner* transformation: the operators $a_x^\pm = \prod_{y=1}^{x-1}(-\sigma_y^3)\sigma_x^\pm$ are a set of anticommuting fermionic operators and , if $\sigma_x^\pm = (\sigma_x^1 \pm i\sigma_x^2)/2$, we can write

$$\sigma_x^- = e^{-i\pi\sum_{y=1}^{x-1} a_y^+ a_y^-} a_x^- \ , \quad \sigma_x^+ = a_x^+ e^{i\pi\sum_{y=1}^{x-1} a_y^+ a_y^-} \ , \quad \sigma_x^3 = 2a_x^+ a_x^- - 1 \ . \quad (6.18)$$

Hence, if we fix the units so that $J_1 + J_2 = 2$ we get

$$\begin{aligned} H = \ & -\frac{1}{2}\sum_{x=1}^{L-1}[a_x^+ a_{x+1}^- + a_{x+1}^+ a_x^-] - u\sum_{x=1}^{L-1}[a_x^+ a_{x+1}^+ + a_{x+1}^- a_x^-] \quad (6.19) \\ & + h\sum_{x=1}^{L}\left(a_x^+ a_x^- - \frac{1}{2}\right) + \lambda \sum_{1\le x,y\le L} v(x-y)\left(a_x^+ a_x^- - \frac{1}{2}\right)\left(a_y^+ a_y^- - \frac{1}{2}\right) \end{aligned}$$

where $\rho_x = a_x^+ a_x^-$, $u = (J_1 - J_2)/2$; it is possible to choose U_L so that periodic boundary conditions are imposed in (6.19). In this form, the model describes interacting non relativistic 1D fermions on a lattice with a short range interaction and a BCS-like term (in the anisotropic case $J_1 \neq J_2$), and it can be used to describe the properties of the conduction electrons of one-dimensional metals.

If O_x is a local monomial in the S_x^α or $a_x^\pm$ operators, we call $O_{\mathbf{x}} = e^{Hx_0} O_x e^{-Hx_0}$ where $\mathbf{x} = (x, x_0)$; moreover, if $A = O_{\mathbf{x}_1} \cdots O_{\mathbf{x}_n}$, we denote its expectation in the grand canonical ensemble by

$$\langle A \rangle_{L,\beta} = \frac{\mathrm{Tr}[e^{-\beta H}\mathbf{T}(A)]}{Tr[e^{-\beta H}]} \quad (6.20)$$

with $\mathbf{T}$ being the time order product; $\langle A \rangle_{T;L,\beta}$ denotes the corresponding truncated expectation. We will be interested in $< A >_T = \lim_{L,\beta\to\infty} < A >_{T;L,\beta}$.

By Renormalization Group methods it was proved in [12] for small λ, $J_1 = J_2 = 1$ and large $\mathbf{x}$,

$$\langle a^-_{\mathbf{x}} a^+_{\mathbf{0}} \rangle_T \sim g_0(\mathbf{x}) \frac{1 + \lambda f(\lambda)}{(x_0^2 + v_s^2 x^2)^{(\eta/2)}} \ , \tag{6.21}$$

where $f(\lambda)$ is a bounded function, $\eta = a_0 \lambda^2 + O(\lambda^3)$, with $a_0 > 0$, and

$$g_0(\mathbf{x}) = \sum_{\omega = \pm} \frac{e^{i\omega p_F x}}{-i x_0 + \omega v_s x} \ , \tag{6.22}$$

$$v_s = v_F + O(\lambda) \qquad p_F = \cos^{-1}(h + \lambda) + O(\lambda) \qquad v_F = \sin p_F \ . \tag{6.23}$$

From (6.21) we see that the interaction has the effect to change the value of the *Fermi momentum* from $\cos^{-1}(h)$ to p_F, and the *Fermi velocity* from v_F in the non interacting case to v_s, and that the power law decay is changed. It was also proved in [12] that the spin-spin correlation in the direction of the 3-axis (or, equivalently, the fermionic density-density correlation) is given, for large $\mathbf{x}$, by

$$\langle S^{(3)}_{\mathbf{x}} S^{(3)}_{\mathbf{0}} \rangle_T \sim \cos(2 p_F x) \Omega^{3,a}(\mathbf{x}) + \Omega^{3,b}(\mathbf{x}) \ , \tag{6.24}$$

$$\Omega^{3,a}(\mathbf{x}) = \frac{1 + A_1(\mathbf{x})}{2\pi^2 [x^2 + (v_s x_0)^2]^{X_+}} \ , \tag{6.25}$$

$$\Omega^{3,b}(\mathbf{x}) = \frac{1}{2\pi^2 [x^2 + (v_s x_0)^2]} \left\{ \frac{x_0^2 - (x/v_s)^2}{x^2 + (v_s x_0)^2} + A_2(\mathbf{x}) \right\} \ , \tag{6.26}$$

with $|A_1(\mathbf{x})|, |A_2(\mathbf{x})| \le C|\lambda|$ and $X_+ = 1 - a_1 \lambda + O(\lambda^2)$ with

$$a_1 = [\widehat{v}(0) - \widehat{v}(2p_F)]/(\pi \sin p_F) \tag{6.27}$$

Finally the Cooper pair density correlation, that is the correlation of the operator $\rho^c_{\mathbf{x}} = a^+_{\mathbf{x}} a^+_{\mathbf{x}'} + a^-_{\mathbf{x}} a^-_{\mathbf{x}'}$, $\mathbf{x}' = (x+1, x_0)$, behaves as

$$\langle \rho^c_{\mathbf{x}} \rho^c_{\mathbf{0}} \rangle_T \sim \frac{1 + A_3(\mathbf{x})}{2\pi^2 (x^2 + v_s^2 x_0^2)^{X_-}} \ , \tag{6.28}$$

with $X_- = 1 + a_1 \lambda + O(\lambda^2)$, a_1 being the same constant appearing in the first order of X_+. In the case $J_1 \neq J_2$ the correlations decay faster than any power with rate ξ such that $\xi \sim C|J_1 - J_2|^{\bar{\nu}}$ with $\bar{\nu} = 1 + a_1 \lambda + O(\lambda^2)$, a_1 given by (6.27).

The same strategy followed for proving the universal relations in the coupled Ising models (1.6) allows to derive the same relations between the indices appearing in the correlations of the spin chain; again all the indices can be expressed in terms of a single one. There is in this case also an extra relation connecting the indices with the Fermi velocity v_s and the susceptibility, defined as

$$\kappa = \lim_{p \to 0} \widehat{\Omega}(0, p) \tag{6.29}$$

where $\widehat{\Omega}(0,p)$ is the bidimensional Fourier transform of $\langle S^{(3)}_{\mathbf{x}} S^{(3)}_{\mathbf{0}}\rangle_T$. In the fermionic interpretation, $\kappa\rho^{-2}$ is the compressibility (ρ is the fermionic density).

Theorem 6.1. *(Benfatto,Mastropietro [13]) In the model (6.17) for λ small enough the exponents in (6.21),(6.24),(6.28) obey*

$$X_+ X_- = 1 \quad \bar\nu = \frac{1}{2 - X_+^{-1}} \qquad 2\eta = X_+ + X_+^{-1} - 2 \tag{6.30}$$

Moreover the susceptibility κ obeys

$$\kappa = \frac{1}{\pi}\frac{X_+}{v_s} \tag{6.31}$$

The relations (6.30) were conjectured by Luther and Peschel [38], while (6.31) connecting the susceptibility defined in (6.29) with X_+ and v_s was conjectured by Haldane in [26]; it is part, together with (6.30), of the so called *Luttinger liquid conjecture*. Note that, using (6.31) and (6.30), from the knowledge of the suscebtibility and the Fermi velocity, all the exponents can be determined.

In the case of the XYZ model ($J_1 \neq J_2$) the exponent $\bar\nu$ has been computed by Baxter and it has been found, see (10.12.24) of [5], if $\cos\bar\mu = -J_3/J_1 = \lambda$,

$$\bar\nu = \frac{\pi}{2\bar\mu} = 1 + \frac{2\lambda}{\pi} + O(\lambda^2)\ . \tag{6.32}$$

and from (6.30) $X_- = 2(1 - \frac{\bar\mu}{\pi})$; from the Bethe ansatz solution [60] exact expressions for v_s and κ can be obtained,

$$v_s = \frac{\pi}{\bar\mu}\sin\bar\mu \qquad \kappa = [2\pi(\pi/\bar\mu - 1)\sin\bar\mu]^{-1}\ , \tag{6.33}$$

and one can verify that (6.31) is verified in the special case of the XYZ model. Either κ, K, v_s depend in general on the magnetic field h and the specific form of the interaction $\widehat{v}(k)$, see (6.27). While there is no hope to understand how the above explicit exact formulas (6.32), (6.33) change when $h \neq 0$ and for generic $v(k)$ (we know them only as series), our theorem says that (6.31) is still true. Its proof is more complex than the proof of the relations between the exponents (which is similar to the one discussed in §5); contrary to the exponents, κ and v_s are not function of $\lambda_{-\infty}$ only (that is, they are not universal in this sense) but their product is nevertheless universal, see [13].

7. Conclusions and Open Problems

The validity of a number of universal relations between exponents and other quantities in a wide class models, including solvable and *not solvable* models, has

been established. The critical exponents are model dependent but satisfy model independent relations, so allowing, for instance, to express all the exponents in terms of a single one, or in terms of other quantities like the susceptibilities. Our results provide one of the very few cases in which the *universality* hypothesis, of so wide use in statistical physics, can be *rigorously* verified.

Of course, the issue of universality even in the class of models we have considered still presents a number of open and challenging problems to mathematical physics. A first one is to prove Theorems 4.1 or 5.1 directly in the spin variables; it is somewhat surprising that one has to pass to the Grassmann integral representation to prove them, and one can wonder why similar results cannot be proved directly in the "natural" spin representation. Also, in Theorems 4.1. or 5.1 the coupling λ must be chosen very small, and the extension of the proof to the optimal value of λ (which is expected anyway to be finite from the exact solutions) would be very important.

A fundamental open problem for spin models like (2.1) or (2.7) is the computation of the spin-spin correlations $\langle \sigma_{\mathbf{x}} \sigma_{\mathbf{y}} \rangle$. Already in the solvable Ising case the analysis of such correlations is very tricky: it is based on an asymptotic analysis of a Toeplitz determinant or, alternatively, on the derivation of highly non trivial non linear finite difference equations, whose scaling limit is related to the third Painleve' equation [44]. In terms of the Grassmann integral representation in §3, the spin correlations are expressed by non gaussian Grassmann integrals of exponentials of quadratic forms summed over a line; their analyis is therefore much more difficult than the energy or crossover correlations, in which one has to integrate a monomial. There are up to now no results for the critical exponents of the spin-spin correlation either in the case of a single perturbed Ising model (2.1) or in the case of a couple of interacting Ising models (2.7); it is believed, see [29], that the analysis of the corresponding Grassmann integrals is related to a phenomenon known in quantum Field Theory as *bosonization*, analyzed rigorously in [8], and perhaps some progress could be done using this idea. Other interesting problems for spin models like (2.7), just to mention a few, are: the analysis of the correlations of (2.7) directly at the critical point; the proof of the conjectures about universal relations for the critical amplitudes; the analysis of four or more interacting Ising models.

Regarding the analysis of quantum Hamiltonians like (6.17) or (6.19), an important open problem is the determination of critical exponents and universality relations for the real time (or *dynamic*) correlations of quantum Hamiltonians like (6.17) or (6.19), namely (6.20) with x_0 replaced with ix_0;such an issue is important for the understanding of transport properties of spin chains. Another important problem regards the proof of universal relations in the one dimensional Hubbard model, that is the model (2.1) in which the fermionic operators have an extra index for the spin: in such a case the phenomenon of spin-charge separation is expected, which seems important for the understanding of high T_c superconductivity, and relations similar to (6.31) are believed to be true for the charge or spin susceptibilities.

References

[1] Adler S. L., Bardeen W.A.: Absence of Higher-Order Corrections in the Anomalous Axial-Vector Divergence Equation. Phys. Rev. 182, 1517–1536 (1969)

[2] Aizenmann A.: Geometric Analysis of ϕ^4 Fields and Ising Models. Comm.Math.Phys 86, 1–48 (1982)

[3] Battle G., Federbush P.: A phase cell cluster expansion for Euclidean field theory. Ann.Phys. 142,95 (1982); Brydges D.: A short course on Cluster Expansions, Les Houches 1984, K. Osterwalder, R. Stora eds., North Holland Press, (1986).

[4] Baxter R.J.: Eight-vertex model in lattice statistics. Phys. Rev. Lett. 26, 832–833, (1971).

[5] Baxter R.J.: *Exactly solved models in statistical mechanics.* Academic Press, Inc. London, (1989).

[6] Benfatto G., Gallavotti G: Perturbation Theory and the Fermi surface in a quantum liquid. A general quasi-particle formalism and one dimensional systems. Jour. Stat. Phys. 59, 541–664 (1990).

[7] Benfatto G., Gallavotti G, Procacci, A, Scoppola B: Beta Functions and Schwinger Functions for a Many Fermions System in One Dimension. Comm. Math. Phys. 160, 93–171 (1994).

[8] Benfatto G., Falco P, Mastropietro V: Massless sine-Gordon and massive Thirring models: proof of Coleman's equivalence. Comm. Math. Phys. 285,2, 713–762 (2009)

[9] Benfatto G., Falco P., Mastropietro V.: Functional Integral Construction of the Massive Thirring model: Verification of Axioms and Massless Limit. Comm. Math. Phys. 273, 67–118, (2007).

[10] Benfatto G., Falco P, Mastropietro V: Extended scaling relations for planar lattice models. Comm. Math. Phys. 292 ,2, 569–605. (2009)

[11] Benfatto G., Falco P, Mastropietro V: Universal Relations for Non Solvable Statistical Models. Phys. Rev. Lett. 104 075701 (2010).

[12] Benfatto G., Mastropietro V.: Renormalization group, hidden symmetries and approximate Ward identities in the XYZ model. Rev. Math. Phys. 13, 1323–1435, (2001).

[13] Benfatto G., Mastropietro V.: Universal relations in non solvable quantum spin chains. J. Stat. Phys. 138, 6,1084 (2010).

[14] Benfatto G., Mastropietro V.: Ward identities and chiral anomaly in the Luttinger liquid. Comm. Math. Phys. 258, 609–655, (2005).

[15] Caianiello E.R.: Number of Feynman graphs and convergence. Nuovo Cimento 10, 1634, (1960).

[16] Feldman J., Magnen J., Rivasseau V, Sénéor R.: Massive Gross-Neveu model: a renormalizable theory in two dimensions. Comm. Math. Phys. 103, 67–103 (1986).

[17] Feldman J, Knoerrer H, Trubowitz E. *Fermionic functional integrals and the Renormalization Group.* CRM series, vol. 16, AMS (2002).

[18] Fröhlich J. On the triviality of $\lambda\phi^4$ theories and the approach to the critical point in $d > 4$ dimensions. Nucl. Phys. B 200, 2,281 (1982).

[19] Gallavotti, G.: Twistless KAM tori. Comm. Math. Phys. 164, 145–156 (1994).

[20] Gallavotti,G. Renormalization theory and ultraviolet stability for scalar fields via renormalization group methods. Rev. Mod. Phys. 57, 471–562 (1985).

[21] Gallavotti,G, Nicolo'.: Renormalization theory in four-dimensional scalar fields. Comm. Math. Phys. 100 (1985), no. 4, 545–590.

[22] Gawedzki K., Kupiainen A.: Gross–Neveu model through convergent perturbation expansions. Comm.Math.Phys. 102, 1–30 (1985).

[23] Giuliani A., Mastropietro V.: Anomalous critical exponents in the anisotropic Ashkin-Teller model. Phys. Rev. Lett. **93**, 190603–07, (2004).

[24] Giuliani A., Mastropietro V.: Anomalous universality in the anisotropic Ashkin-Teller model. Comm. Math. Phys. **256**, 681–725, (2005).

[25] Glimm J, Jaffe A. *Quantum Physics: A Functional Integral Point of View* Springer-Verlag (1987)

[26] Haldane D.M.: General relation of correlarion exponents and application to the anisotropic S=1/2 Heisenberg chain. Phys. Rev. Lett. 45, 1358–1362, (1980).

[27] Hocken R, Moldover M.: Ising critical exponents in real fluids: an experiment. Phys Rev Lett 37, 29–32 (1976)

[28] Hurst C.A.: New approach to the Ising model. J. Math. Phys. 7, 305–310 (1966).

[29] Itzykson C., Drouffe J. *Statistical field theory.* Cam. Un. Press (1989)

[30] Johnson K.: Solutions of the equations for the Green functions of a two dimensional relativistic quantum field theory. Nuovo Cimento 20, 773–790, (1961).

[31] Kadanoff L.P.: Connections between the Critical Behavior of the Planar Model and That of the Eight-Vertex Model. Phys. Rev. Lett. 39, 903–905, (1977).

[32] Kadanoff L.P., Brown A.C.:Correlation functions on the critical lines of the Baxter and Ashkin-Teller models. Ann. Phys. 121, 318–345, (1979).

[33] Kadanoff L.P., Wegner F.J.: Connections between the Critical Behavior of the Planar Model and That of the Eight-Vertex Model. Phys. Rev. B 4, 3989–3993, (1971).

[34] Kasteleyn P.W.: Dimer statistics and phase transition. J. Math. Phys. 4, 287, (1963).

[35] A. Lesniewski: Effective action for the Yukawa$_2$ quantum field theory. Comm. Math. Phys. 108, 437–467, (1987).

[36] Lieb E.: Exact solution of the problem of the entropy of two dimensional ice. Phys. Rev. Lett 18, 692, (1967)

[37] J. A. Lipa, D. R. Swanson, et. al.. Heat Capacity and Thermal Relaxation of Bulk Helium very near the Lambda Point. Phys. Rev. Lett. 76, 944 (1996).

[38] Luther A., Peschel I.: Calculation of critical exponents from quantum field theory in one dimension. Phys. Rev. B, 12, 3908–3917, (1975).

[39] Mastropietro V.: Non-universality in Ising models with four spin interaction. J. Stat. Phys. 111, 201–259, (2003).

[40] Mastropietro V.: Ising models with four spin interaction at criticality. Comm. Math. Phys. 244, 595–64 (2004).

[41] Mastropietro V.: Nonperturbative Adler-Bardeen theorem. J. Math. Phys 48, 022302, (2007).

[42] Mastropietro V.: *Non-perturbative Renormalization.* World Scientific, (2008).

[43] D. Mattis, E. Lieb. Exact solution of a many fermion system and its associated boson field. J. Math. Phys. 6, 304–312 (1965).

[44] McCoy B., Wu T.: *The two dimensional Ising model*. Harward Univ. Press (1973).

[45] den Nijs M.P.M.: Derivation of extended scaling relations between critical exponents in two dimensional models from the one dimensional Luttinger model. Phys. Rev. B 23, 6111–6125, (1981).

[46] Onsager L.: Critical statistics. A two dimensional Ising model with an order-disorder trasition. Phys. Rev., 65, 117–149 (1944)

[47] Polchinski,J.: Renormalization and effective Lagrangians. Nucl. Phys. B 231 , 269 (1984)

[48] Pruisken A.M.M. Brown A.C.: Universality for the critical lines of the eight vertex, Ashkin-Teller and Gaussian models. Phys. Rev. B 23, 1459–1468, (1981).

[49] Pruisken A.M.M., L.P. Kadanoff L.P.: Marginality, universality, and expansion techniques for critical lines in two dimensions. Phys. Rev. B 22 5154 (1980).

[50] Ruelle D. *Statistical Mechanics*, Benjiamin, New York (1969)

[51] Samuel S.: The use of anticommuting variable integrals in statistical mechanics. J. Math. Phys. 21, 2806, (1980).

[52] Schultz T.D., Mattis D.C., Lieb E.: Two dimensional Ising model as soluble problem of many fermions. Rev. Mod. Phys. 36, 856–871, (1964).

[53] Smirnov S.: Towards conformal invariance of 2D lattice models. Proceedings Madrid ICM, Europ. Math. Soc, (2006)

[54] Spencer T.: A mathematical approach to universality in two dimensions. Physica A 279, 250–259, (2000); Pinson H., Spencer T.: Universality in 2D Critical Ising model. Preprint

[55] Spohn H.: Bosonization, vicinal surfaces, and hydrodynamic fluctuation theory. Phys. Rev. E 60, 6411 (1999).

[56] Sutherland,W.: Exact solution of a two dimensional model for hydrogen-bonded crystals. Phys. Rev. Lett 19, 103, (1967)

[57] Sutherland D.: Two-Dimensional Hydrogen Bonded Crystals without the Ice Rule. J.Math.Phys 11, 3183–86 (1970)

[58] Widom B.: Equation of state in the neighborhood of the critical point. J.Chem.Phys. 43, 3892–7 and 3898–905 (1965)

[59] Wilson K.G.: The Renormalization Group and the critical phenomena. Rev.Mod.Phys. 55,583–600 (1983)

[60] Yang C.N.,Yang C.P.: Ground state energy of a Heisenber-Ising lattice. Phys. Rev. 147, 303–306 (1966)

[61] Zamolodchikov A.B., Zamolodchikov Al. B.: Conformal field theory and 2D critical phenomena. Soviet Scientific Reviews A 10, 269, (1989).

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Weak Solutions to the Navier-Stokes Equations with Bounded Scale-invariant Quantities

Gregory A. Seregin*

Abstract

The main assumption of the so-called ε-regularity theory of suitable weak solutions to the Navier-Stokes equations is uniform smallness of certain scale-invariant quantities, which rules out singularities. One of the best results of ε-regularity is the famous Caffarelli-Kohn-Nirenberg theorem. Our goal is to understand what happens if the assumption on smallness of scale-invariant quantities is replaced with their uniform boundedness. The latter makes it possible to use blow-up technique and reduce the local regularity problem to the question of existence or non-existence of "non-trivial" ancient (backward) solutions to the Navier-Stokes equations. There are at least two potential scenarios: the classical Liouville type problem for mild bounded ancient solutions and backward uniqueness for the Navier-Stokes equations. In this survey, we discuss sufficient conditions implying non-existence of "non-trivial" solutions and the corresponding sufficient conditions ensuring local regularity of original weak solutions.

Mathematics Subject Classification (2010). Primary 35Q30, Secondary 76D05.

Keywords. Navier-Stokes equations, regularity, weak Leray-Hopf solutions, suitable weak solutions, ancient solutions.

1. Introduction

One of the main problems of the mathematical hydrodynamics can be formulated as follows. Is the Cauchy problem, describing the flow of viscous incompressible fluids, *globally well-posed*? In other words, given a smooth divergence

*OxPDE, Mathematical Institute, University of Oxford, UK[1].
E-mail: gaseregin@googlemail.com.
[1]on leave from POMI, St.Petersburg, Russia

free velocity field a compactly supported in $\mathbb{R}^3$, does the classical Navier-Stokes system

$$\partial_t v(x,t) + v(x,t)\cdot\nabla v(x,t) - \nu\Delta v(x,t) + \nabla q(x,t) = 0,$$
$$\operatorname{div} v(x,t) = 0 \tag{1.1}$$

have a unique solution subject to the initial condition

$$v(x,0) = a(x), \qquad x \in \mathbb{R}^3, \tag{1.2}$$

which is defined globally for all $x \in \mathbb{R}^3$ and for all $0 < t < +\infty$? Here, as usual, v and q stand for the velocity field and for the pressure field, respectively. In this paper, we are not going to study the very important issue how solutions depend on the viscosity ν and let it equal to 1.

There are two different approaches to attack the above problem. In one of them, the Cauchy problem (1.1) and (1.2) can be reformulated as an integral equation by removing the non-linear term to the right hand side of the first equation in (1.1), by applying Leray's projector P to both sides of it, and by inverting then the linear part. As a result, the following equation with respect to a function v of time t with values in a Banach space appears

$$v(t) = S(t)a - \int_0^t S(t-\tau)P(v(\tau)\cdot\nabla v(\tau))d\tau.$$

Here, $S(t)$ is the solution operator of the Cauchy problem for the heat equation. Any solution to the above integral equation is called a *mild* solution to the Cauchy problem (1.1) and (1.2). Existence and uniqueness of mild solutions can be proved with the help of contraction mappings. For history, details, and references, we recommend papers [14], [11], [2], and [18]. This approach is quite effective for proving local well-posedness for a wide range of initial data.

Another method gives *energy* solutions called nowadays *weak Leray-Hopf* solutions. They have been introduced by J. Leray in his pioneering paper [22] for the Cauchy problem and in a sense by E. Hopf in [13] for initial boundary value problems. The modern definition of weak solutions can be found, for example, in [19] and includes the following ingredients:

(i) the velocity field v has finite kinetic energy and finite dissipation;

(ii) the Navier-Stokes system is satisfied in the sense of distributions with divergence free test functions;

(iii) the velocity field v is a continuous function of time t with values in the space L_2 equipped with the weak topology;

(iv) the initial data are fulfilled in the strong L_2-sense;

(v) the velocity field v satisfies the global energy inequality for all possible values of t.

Not all the above properties are independent each of other. Choosing divergence free test functions in (ii), we exclude the pressure field from the definition completely.

So, we have a global weak Leray-Hopf solution to the classical problem but we do not know whether it is unique or not. However, as it has been observed and proved by J. Leray in [22], *any smooth solution to the classical Cauchy problem is unique in the class of weak solutions.* In other words, the problem of uniqueness of weak solutions can be posed as a more particular problem of their smoothness. The latter has been proposed as one of the seven Millennium problems in [10].

By definition, the space-time point $z = (x, t)$ is called a *regular point of the velocity field* v if v is of class L_∞ in a parabolic vicinity with the center at z. The first moment of time T when singularities occur is called a *blowup time*. Further smoothing in a neighborhood of a regular point is straightforward and a simple consequence of the linear theory.

Our approach to the regularity problem is quite typical for the classical theory of partial differential equations, namely, we are going to study smoothness of weak solutions locally in space-time. Now, let us state the local regularity problem for the Navier-Stokes equations rigorously.

Consider the Navier-Stokes system (1.1) in a canonical domain, say, in the unit parabolic cylinder Q being the Cartesian product of the unit ball B of $\mathbb{R}^3$ with the center at the origin and the time interval $]-1, 0[$. More general cases can be reduced to the canonical one with the help of the space-time shift and the Navier-Stokes scaling

$$\begin{aligned} v^\lambda(y, s) &= \lambda v(x_0 + \lambda y, t_0 + \lambda^2 s), \\ q^\lambda(y, s) &= \lambda^2 q(x_0 + \lambda y, t_0 + \lambda^2 s). \end{aligned} \tag{1.3}$$

Our question is as follows. What are the weakest assumptions on v and q that provide regularity of v at the origin $z = (x, t) = 0$? Ideally, they should be fulfilled for energy solutions but this is unknown and might be not necessary true.

Nowadays, it is well understood that the main object of the local regularity theory of the Navier-Stokes equations is the so-called suitable weak solutions. They were introduced in the middle 70s by V. Scheffer in the series of papers, see for example [25] and [26], where the importance of solutions satisfying the local energy inequality was pointed out and exploited. In the early 80s, the essential contribution to understanding suitable weak solutions in the context of the local regularity theory was made in the celebrated paper [1] by L. Caffarelli, R.-V. Kohn, and L. Nirenberg. However, in this survey, we are going to accept a more

particular but very much convenient version of the definition of suitable weak solutions given by F.-H. Lin in [23].

Definition 1.1. *Functions $v \in L_{2,\infty}(Q) \cap W_2^{1,0}(Q)$ and $q \in L_{\frac{3}{2}}(Q)$ are said to be a suitable weak solution to the Navier-Stokes equations in Q if they satisfy (1.1) in Q in the sense of distributions and, for a.a. $t \in]-1,0[$, the local energy inequality*

$$\int\limits_B \varphi(x,t)|v(x,t)|^2 dx + 2\int\limits_{-1}^{t}\int\limits_B \varphi|\nabla v|^2 dxdt' \le \int\limits_{-1}^{t}\int\limits_B \Big(|v|^2(\partial_t\varphi + \Delta\varphi) + v\cdot\nabla\varphi(|v|^2+2q)\Big)dxdt'$$

holds for any non-negative test function $\varphi \in C_0^\infty(B\times]-1,1[)$.

Here, the following notation for mixed Lebesgue and Sobolev spaces is used: $L_{m,n}(Q) = L_n(-1,0;L_m(B))$, $L_{m,m} = L_m$, $W^{1,0}_{m,n}(Q) = \{v, \nabla v \in L_{m,n}(Q)\}$, and $W^{1,0}_{m,m} = W^{1,0}_m$.

It is not so difficult to show that among weak Leray-Hopf solutions to a given Cauchy problem (1.1) and (1.2), there is at least one with the following property. For any point $z_0 = (x_0, t_0)$ with $t_0 > 0$, this solution v, together with the *associated pressure* q, satisfies all the requirements to be a suitable weak solution to the Navier-Stokes equations in $z_0 + Q(R)$ for any $R > 0$ subject to the restriction $t_0 - R^2 > 0$. By $Q(R)$, we denote a parabolic ball (cylinder) of $\mathbb{R}^3 \times \mathbb{R}$ with radius R centered at the origin, i.e., $Q(R) = B(R)\times]-R^2, 0[$ and $B(R) \subset \mathbb{R}^3$ is a ball of radius R with the center at the origin.

Another relatively well-understood thing is the role of quantities invariant with respect to the Navier-Stokes scaling (1.3) with $x_0 = 0$ and $t_0 = 0$. By the definition, such quantities are defined on parabolic balls $Q(r)$ and have the property $F(v,q;r) = F(v^\lambda, q^\lambda; r/\lambda)$.

Now, we are in a position to explain the so-called ε-regularity theory for suitable weak solutions to the Navier-Stokes equations. There are two types of statements in it and the first one essentially proved in [1], see also [24], and reads:

Suppose that v and q are a suitable weak solution to the Navier-Stokes equations in Q. There exist universal positive constants ε and c_k, $k = 0,1,2,...$ such that if $F(v,q;1) < \varepsilon$ then $|\nabla^k v(0)| < c_k$, $k = 0,1,2,....$ Moreover, the function $z \mapsto \nabla^k v(z)$ is Hölder continuous (relative to the usual parabolic metric) with any exponent less $1/3$ in the closure of $Q(1/2)$.

Here, it is an important example of such kind of quantities:

$$F(v,q;r)=\frac{1}{r^2}\int\limits_{Q(r)}\Big(|v|^3+|q|^{\frac{3}{2}}\Big)dz.$$

The limited smoothing in time cannot be improved. This can be easily seen from Serrin's example

$$\begin{aligned}v(x,t)&=C(t)\nabla h(x),\\ q(x,t)&=-C'(t)h(x)-1/2C(t)|\nabla h(x)|^2,\end{aligned}$$

in which h is a harmonic function of x and C is a given function of t.

In the other type of statements, it is supposed that our quantity F is independent of the pressure q:

Let v and q be a suitable weak solution in Q. There exist a universal positive constant ε with the property: if $\sup_{0<r<1}F(v;r)<\varepsilon$ then $z=0$ is a regular point. Moreover, for any $k=0,1,2,...$, the function $z\mapsto\nabla^k v(z)$ is Hölder continuous with any exponent less $1/3$ in the closure of $Q(r)$ for some positive r.

Dependence on the pressure in the above statement is hidden. In fact, the radius r is determined by the $L_{\frac{3}{2}}$-norm of the pressure over the whole parabolic cylinder Q.

To illustrate the second statement, let us consider several examples. In the first one, we deal with the Ladyzhenskaya-Prodi-Serrin type quantities

$$F(v;r)=M_{s,l}(v;r)=\|v\|^l_{s,l,Q(r)}=\int\limits_{-r^2}^{0}\left(\int\limits_{B(r)}|v|^s dx\right)^{\frac{l}{s}}dt$$

provided

$$\frac{3}{s}+\frac{2}{l}=1$$

and $s\geq 3$. Local regularity results connected with those quantities have been proved partially by J. Serrin in [34] and then by M. Struwe in [35] for the velocity field v having finite energy even with no assumption on the pressure. However, in such a case, we loose Hölder continuity, see the above example.

The second kind of quantities will be called scaled energy quantities. Let us list some of them:

$$A(v;r) = \sup_{-r^2<t<0} \frac{1}{r} \int\limits_{B(r)} |v(x,t)|^2 dx,$$
$$C(v;r) = \frac{1}{r^2} \int\limits_{Q(r)} |v|^3 dz,$$
$$E(v;r) = \frac{1}{r} \int\limits_{Q(r)} |\nabla v|^2 dz,$$
$$D(q;r) = \frac{1}{r^2} \int\limits_{Q(r)} |q|^{\frac{3}{2}} dz.$$

For more examples of scaled energy quantities, we refer to the paper [12]. It is interesting to note that the second statement applied to the scaled dissipation E is the famous Caffarelli-Kohn-Nirenberg theorem. It gives the best estimate for Hausdorff's dimension of the singular set for a class of weak Leray-Hopf solutions to the Cauchy problem. A sort of logarithmic improvement of the latter result is explained in [6]. A certain generalization of the Caffarelli-Kohn-Nirenber theorem itself has been proved in [28] and is formulated as follows.

Proposition 1.2. *Let v and q be a suitable weak solution to the Navier-Stokes equations in Q. Given $M > 0$, there exists a positive number $\varepsilon(M)$ having the property: if two inequalities $\limsup_{r\to 0} E(r) < M$ and $\liminf_{r\to 0} E(r) < \varepsilon(M)$ hold, then $z = 0$ is a regular point of v.*

Typical examples of the third group of quantities invariant to the Navier-Stokes scaling are:

$$G_1(v;r) = \sup_{z=(x,t)\in Q(r)} |x||v(z)|,$$
$$G_2(v;r) = \sup_{z=(x,t)\in Q(r)} \sqrt{-t}|v(z)|.$$

A proof of the corresponding statements has been presented in [32], see also [36], [16], and [5] for similar results.

The question we are interested in is *what happens if we drop the condition on smallness of scale-invariant quantities, assuming their uniform boundedness only, i.e,* $\sup_{0<r<1} F(v,r) < +\infty$. For Ladyzheskaya-Prodi-Serrin type quantities with $s > 3$, the answer is still positive, i.e., $z = 0$ is a regular point. It follows from scale-invariance and the fact that the assumption $M_{s,l}(v;1) = \sup_{0<r<1} M_{s,l}(v;r) < +\infty$ implies $M_{s,l}(v;r) \to 0$ as $r \to 0$ if

$s > 3$. Although in the marginal case $s = 3$ and $l = +\infty$ the answer reamins positive, the known proof is more complicated and will be outlined later.

In this review, we shall discuss various approaches to the problem in question. Before going into details, let us recall certain definitions and make some general remarks about relationships between some scale-invariant quantities. Boundedness of $\sup_{0<r<1} G_2(v;r) = G_2(v,1) = G_{20} < +\infty$ can be rewritten in the form

$$|v(z)| \le \frac{G_{20}}{\sqrt{-t}}$$

for all $z = (x,t) \in Q$. If v satisfies the above inequality and $z = 0$ is still a singular point of v, we say that a *singularity of Type I* or *Type I blowup* takes place at $t = 0$. All other singularities are of Type II. The main feature of Type I singularities is that they have the same rate as potential self-similar solutions. The important properties connected with possible singularities of Type I have been proved in [31] and are as follows.

Proposition 1.3. *Let functions v and q be a suitable weak solution to the Navier-Stokes equations in Q.*

(i) If $\min\{G_1(v;1), G_2(v;1)\} < +\infty$, *then*

$$g = \sup_{0<r<1} \{A(v;r) + C(v;r) + D(q;r) + E(v;r)\} < +\infty.$$

(ii) If

$$g' = \min\{\sup_{0<r<1} A(v;r), \sup_{0<r<1} C(v;r), \sup_{0<r<1} E(v;r)\} < +\infty,$$

then $g < +\infty$.

This proposition admits many obvious generalizations.

2. Blowup Techniques, Bounded Ancient Solutions

In this section, we always assume that $z = 0$ is a singular point. Making use of the space-time shift and the Navier-Stokes scaling, we can reduce the general problem of local regularity to a particular one that in a sense mimics the first time singularity.

Proposition 2.1. *Let v and q be a suitable weak solution to the Navier-Stokes equations in Q and $z = 0$ be a singular point of v. There exist two functions $\widetilde{v}$ and $\widetilde{q}$ having the following properties:*

(i) $\widetilde{v} \in L_3(Q)$ and $\widetilde{q} \in L_{\frac{3}{2}}(Q)$ obey the Navier-Stokes equations in Q in the sense of distributions;

(ii) $\widetilde{v} \in L_\infty(B\times]-1,-a^2[)$ *for all* $a \in]0,1[$*;*

(iii) there exists a number $0 < r_1 < 1$ *such that* $\widetilde{v} \in L_\infty(\{(x,t) : r_1 < |x| < 1,\ -1 < t < 0\})$.

Moreover, functions $\widetilde{v}$ *and* $\widetilde{q}$ *are obtained from* v *and* q *with the help of the space-times shift and the Navier-Stokes scaling and the origin remains to be a singular point of* $\widetilde{v}$.

The proof of Proposition 2.1 is essentially based on the application of the Caffarelli-Kohn-Nirenberg theorem and given in the paper [31]. In what follows, it is always deemed that such a replacement of v and q with $\widetilde{v}$ and $\widetilde{q}$ has been already made. Coming back to the original notation, we assume that functions v and q satisfy all the properties listed in Proposition 2.1 and $z = 0$ is a singular point of v.

One of the most powerful methods to study potential singularities is a blowup technique based on the Navier-Stokes scaling

$$u^{(k)}(y,s) = \lambda_k v(x,t), \qquad p^{(k)}(y,s) = \lambda_k^2 q(x,t)$$

with

$$x = x^{(k)} + \lambda_k y, \qquad x = t_k + \lambda_k^2 s,$$

where $x^{(k)} \in \mathbb{R}^3$, $-1 < t_k \le 0$, and $\lambda_k > 0$ are parameters of the scaling and $\lambda_k \to 0$ as $k \to +\infty$. It is supposed that functions v and q are extended by zero to the whole $\mathbb{R}^3 \times \mathbb{R}$. A particular selection of scaling parameters $x^{(k)}$, t_k, and λ_k depends upon a problem under consideration.

Now, our goal is to describe a universal method that makes it possible to reformulate the local regularity problem as a classical Liouville type problem for the Navier-Stokes equations. To see how things work, let us introduce the function

$$M(t) = \sup_{-1<\tau\le t} \|v(\cdot,\tau)\|_{\infty,\overline{B}(r_1)}.$$

It tends to infinity as time t goes to zero from the left since the origin is a singular point of v. Thanks to the obvious properties of the function M, one can choose parameters of the scaling in a particular way letting $\lambda_k = 1/M_k$, where a sequence M_k is defined as

$$M_k = \|v(,t_k)\|_{\infty,\overline{B}(r_1)} = |v(x^{(k)},t_k)|.$$

Before discussing what happens if k tends to infinity, let us introduce a subclass of bounded ancient (backward) solutions playing an important role in the regularity theory of the Navier-Stokes equations.

Definition 2.2. *A bounded vector field u, defined on $\mathbb{R}^3\times]-\infty,0[$, is called a mild bounded ancient solution to the Navier-Stokes equation if there exists a function p in $L_\infty(-\infty,0;BMO(\mathbb{R}^3))$ such that u and p satisfy the Navier-Stokes system*

$$\begin{aligned}\partial_t u + \operatorname{div} u\otimes u - \Delta u + \nabla p &= 0,\\ \operatorname{div} u &= 0\end{aligned}$$

in $\mathbb{R}^3\times]-\infty,0[$ in the sense of distributions.

The notion of mild bounded ancient solutions has been introduced in [17]. It has been proved there that u has continuous derivatives of any order in both spatial and time variables. Actually, the definition accepted in the present paper is different but equivalent to the one given in [17]. Here, we follow [31].

The statement below that has been proved in [31] shows how mild bounded ancient solutions appear in the regularity theory of the Navier-Stokes equations.

Proposition 2.3. *There exist a subsequence of $u^{(k)}$ (still denoted by $u^{(k)}$) and a mild bounded ancient solution u such that, for any $a>0$, the sequence $u^{(k)}$ converges uniformly to u on the closure of the set $Q(a)=B(a)\times]-a^2,0[$. The function u has the additional properties: $|u|\le 1$ in $\mathbb{R}^3\times]-\infty,0[$ and $|u(0)|=1$.*

Let us demonstrate how this method works in the simplest case of the regular Ladyzhenskaya-Prodi-Serrin quantity $M_{5,5}$. Suppose that

$$M_{5,5}(v,1)=\sup_{0<r<1} M_{5,5}(v,r)<+\infty.$$

By the scale-invariance and by the pressure equation, we may assume without loss of generality that

$$\int\limits_{-\infty}^{0}\int\limits_{\mathbb{R}^3}(|u|^5+|p|^{\frac{5}{2}})dxdt<+\infty.$$

Given $\varepsilon>0$, we can find $T<0$ such that

$$\int\limits_{-\infty}^{T}\int\limits_{\mathbb{R}^3}(|u|^5+|p|^{\frac{5}{2}})dxdt<\varepsilon.$$

Then, by Hölder inequality, we have

$$\frac{1}{R^2}\int\limits_{t_0-R^2}^{t_0}\int\limits_{B(x_0,R)}(|u|^3+|p|^{\frac{3}{2}})dxdt<c\varepsilon^{\frac{3}{5}}$$

for any $x_0\in\mathbb{R}^3$, any $R>0$, and any $t_0\le T$ with some universal constant c. In turn, the ε-regularity theory ensures the inequality

$$|u(x_0,t_0)|<\frac{c}{R}$$

with another universal constant c. Tending $R\to\infty$, we get $u(\cdot,t)=0$ as $t\le T$. One can repeat more or less the same arguments in order to show that in fact u is identically zero on $\mathbb{R}^3\times]-\infty,0]$, which contradicts non-triviality condition $|u(0)|=1$.

It is worthy to notice that the trivial bounded ancient solution of the form

$$u(x,t)=c(t),\qquad p(x,t)=-c'(t)\cdot x,$$

with arbitrary bounded function $c(t)$, is going to be a mild bounded ancient solution if and only if $c(t)\equiv$ *constant*. As in [31], this allows us to make the following plausible conjecture.

Conjecture *Any mild bounded ancient solution to the Navier-Stokes equations is a constant.*

To explain what consequences of the conjecture could be for regularity theory of the Navier-Stokes equations, let us assume that some "reasonable" scale-invariant quantity for v is "uniformly" bounded, see Introduction for definitions. By this assumption, together with Proposition 1.3, and by the conjecture, any mild bounded ancient solution must be zero. However, by Proposition 2.3, if $z=0$ is a singular point, there must be at least one non-trivial mild bounded ancient solution. So, the origin $z=0$ cannot be a singular point of v. This would be a positive answer to the question raised in the introduction. In particular, according to Proposition 1.3, validity of the conjecture would rule out Type I blowups.

As it has been shown in [17], the conjecture is true at least in two non-trivial cases. One of them is the two-dimensional flow for which regularity of energy solutions is well known, see Ladyzhenskaya's monograph [19]. In the second case, axial symmetry with respect to x_3-axis is assumed and the behavior of solutions far away from the axis of symmetry is supposed to respect the property:

$$\sqrt{x_1^2+x_2^2}|u(x,t)|\le C<+\infty$$

for any $x\in R^3$ and for any $-\infty<t<0$. This result can be exploited to

show that boundedness of g', see definition of g' in Proposition 1.3, implies regularity of axially symmetric solutions with no assumption on the swirl. In the corresponding arguments, the crucial point is that, for axially symmetric solutions, boundedness of g' for v provides the required decay of u, see a proof in [33] or in [31]. A simple consequence of the latter statement is that axially symmetric solutions cannot develop Type I blowups. Indeed, to this end, it is sufficient to apply Proposition 1.3 (i) and get boundedness of g'.

Smoothness of axially symmetric solutions with no swirl is well known due to O. A. Ladyzhenskaya in [20] and M. R. Ukhovskij and V. L. Yudovich in [37], while the absence of Type I blowups with no assumptions on swirl has been established relatively recently, see details in [3], [4], [17], and [31].

It is interesting to notice that for the non-regular Ladyzheskaya-Prodi-Serrin condition (so-called $L_{3,\infty}$-case), we are still not able to prove this conjecture. In the next sections, we shall discuss other ways of constructing ancient solutions to the Navier-Stokes equations in order to solve $L_{3,\infty}$-problem.

3. Backward Uniqueness for Navier-Stokes Equations

In this section, we deal with another subclass of ancient solutions u possessing the following property: *there exists a function p defined on $\mathbb{R}^3\times]-\infty,0[$ such that functions u and p are a suitable weak solution to the Navier-Stokes equations in $\mathbb{R}^3\times]-\infty,0[$, i.e., they are a suitable weak solution on each parabolic ball of the form $Q(a)=B(a)\times]-a^2,0[$ with $<a<+\infty$.* We call u a *local energy ancient* solution. Certainly, mild bounded ancient solutions belong to this subclass.

Local energy ancient solutions can be obtained from a given suitable weak solution v and q defined in Q with the help of the scaling mentioned in the previous section provided boundedness of g' takes place, see the definition of g' in Proposition 1.3.

Proposition 3.1. *Let v and q be a suitable weak solution to the Navier-Stokes equations in Q with $g'<+\infty$ and let $u^{(k)}(y,s)=\lambda_k v(\lambda_k y,\lambda_k^2 s)$ and $p^{(k)}(y,s)=\lambda_k^2 q(\lambda_k y,\lambda_k^2 s)$ with $\lambda_k\to 0$ as $k\to+\infty$. Then there exist subsequences of $u^{(k)}$ and $p^{(k)}$ still denoted by $u^{(k)}$ and $p^{(k)}$ such that, for each $a>0$,*

$$u^{(k)}\to u$$

in $L_3(Q(a))\cap C([-a^2,0];L_{\frac{9}{8}}(B(a)))$ and

$$p^{(k)}\rightharpoonup p$$

in $L_{\frac{3}{2}}(Q(a))$, where u is a local energy ancient solution with the corresponding

pressure p. For them, the scaled energy quantities are uniformly bounded, i.e.,

$$\sup_{0<a<+\infty}\{A(u;a)+C(u;a)+D(p;a)+E(u;a)\}<+\infty.$$

Moreover, if $z=0$ is a singular point of the velocity field v, then

$$\int\limits_{Q(3/4)}|u|^3dz>c \tag{3.1}$$

with a positive universal constant c, i.e., u is not identically equal to zero.

A proof of this proposition and similar facts can be found in [8], [29], [31], and [30]. Let us comment the last statement of Proposition 3.1. Indeed, if $z=0$ is a singular point of v, the ε-regularity theory gives us

$$\frac{1}{r^2}\int\limits_{Q(r)}(|v|^3+|q|^{\frac{3}{2}})dz>\varepsilon>0$$

for all $0<r<1$ and for some universal constant ε. Making the inverse change of variables, we find

$$\frac{1}{a^2}\int\limits_{Q(a)}(|u^{(k)}|^3+|p^{(k)}|^{\frac{3}{2}})dyds=$$
$$\frac{1}{\lambda_k^2a^2}\int\limits_{Q(\lambda_ka)}(|v|^3+|q|^{\frac{3}{2}})dxds>\varepsilon>0$$

for each fixed radius $a>0$ and for sufficiently large natural number k. We cannot simply pass to the limit in the latter identity since it is not clear whether the pressure $p^{(k)}$ converges strongly. This is quite typical issue when working with sequences of weak solutions to the Navier-Stokes equations. In order to treat this case one can split the pressure $p^{(k)}$ into two parts. The first part is completely controlled by the velocity field $u^{(k)}$ while the second one is a harmonic function with respect to the spatial variables. This, together with a certain boundedness of the sequence $p^{(k)}$, implies (3.1). For more details, we recommend papers [29] and [30].

We do not know whether local energy ancient solutions with bounded scaled energy quantities are identically equal to zero. However, there are some interesting cases for which the answer is positive. Let us describe them.

Our additional standing assumption of this section can be interpreted as a restriction on the blowup profile of v and has the form

$$\frac{1}{r^{\frac{15}{8}}}\int\limits_{B(r)}|v(x,0)|^{\frac{9}{8}}dx\to 0 \tag{3.2}$$

as $r \to 0$. The most important consequence of (3.2) is that

$$u(\cdot, 0) = 0, \tag{3.3}$$

where u is a local energy ancient solution generated by the scaling of Proposition 3.1. Indeed, for any $a > 0$, we have

$$\frac{1}{a^{\frac{15}{8}}} \int\limits_{B(a)} |u(y,0)|^{\frac{9}{8}} dy \leq$$
$$c\frac{1}{a^{\frac{15}{8}}} \int\limits_{B(a)} |u(y,0) - u^{(k)}(y,0)|^{\frac{9}{8}} dy + c\frac{1}{a^{\frac{15}{8}}} \int\limits_{B(a)} |u^{(k)}(y,0)|^{\frac{9}{8}} dy =$$
$$\alpha_k(a) + c\frac{1}{(\lambda_k a)^{\frac{15}{8}}} \int\limits_{B(\lambda_k a)} |v(x,0)|^{\frac{9}{8}} dx.$$

Now, by Proposition 3.1 and by (3.2), the right hand side of the latter inequality tends to zero and this completes the proof of (3.3).

In a view of (3.3), one could expect that our local energy ancient solution is identically equal to zero. We call this phenomenon a backward uniqueness for the Navier-Stokes equations. So, if the backward uniqueness takes place or at least our ancient solution is zero on the time interval $]-3/4, 0[$, then (3.1) cannot be true and thus, by Proposition 3.1, the origin $z = 0$ is not a singular point of the velocity field v.

The crucial point for understanding the backward uniqueness for the Navier-Stokes equations is a similar phenomenon for the heat operator with lower order terms. The corresponding statement for the partial differential inequality involving the backward heat operator with lower order terms has been proved in [8] and reads:

Theorem 3.2. *Assume that we are given a function ω defined on $\mathbb{R}^n_+ \times]0,1[$, where $\mathbb{R}^n_+ = \{x = (x_i) \in \mathbb{R}^n,\ x_n > 0\}$. Suppose further that they have the properties:*

ω and the generalized derivatives $\nabla\omega$, $\partial_t\omega$, and $\nabla^2\omega$ are square integrable over any bounded subdomain of $\mathbb{R}^n_+ \times]0,1[$;

$$|\partial_t\omega + \Delta\omega| \leq c(|\omega| + |\nabla\omega|) \tag{3.4}$$

on $\mathbb{R}^n_+ \times]0,1[$ with a positive constant c;

$$|\omega(x,t)| \leq \exp\{M|x|^2\} \tag{3.5}$$

for all $x \in \mathbb{R}^n_+$, for all $0 < t < 1$, and for some $M > 0$;

$$\omega(x,0) = 0 \tag{3.6}$$

for all $x \in \mathbb{R}^n_+$.

Then ω is identically zero in $\mathbb{R}^n_+\times]0,1[$.

The interesting feature of Theorem 3.2 is that there has been made no assumption on ω on the boundary $x_n = 0$. In order to prove the theorem, two Carleman's inequalities have been established, see details in [8] and [9]. For the further improvements of the above backward uniqueness result, we refer to the interesting paper [7].

Theorem 3.2 clearly indicates what one should add to (3.3) in order to get the backward uniqueness for ancient solutions to the Navier-Stokes equations. Apparently, we need more regularity for sufficiently large x and a right decay at infinity. One can hope then to apply Theorem 3.2 to the vorticity equation

$$\partial_t\omega - \Delta\omega = \omega\cdot\nabla u - u\cdot\nabla\omega, \qquad \omega = \nabla\wedge u,$$

which could be interpreted as a perturbation of the heat equation by lower order terms. To make it possible, it is sufficient to show boundedness of u and ∇u outside of the Cartesian product of some spatial ball and some time interval. The most of the rest of the paper will be devoted to description of various situations for which it is really true.

Let us assume that

$$|u(x,t)| + |\nabla u(x,t)| \le c < +\infty \tag{3.7}$$

for all $|x| > R$, for all $-1 < t < 0$, and for some constant c and try to figure out what follows from (3.7). It is not difficult to see that (3.3) and (3.7) implies (3.6) and (3.4), (3.5), respectively. At last, the linear theory ensures the validity of first condition in Theorem 3.2, see details in [27]. So, Theorem 3.2 is applicable and by it, $\omega(x,t) = 0$ for all $|x| > R$ and for $-1 < t < 0$. Using unique continuation across spatial boundaries, see, for instance, [8], we deduce $\omega(x,t) = \nabla\wedge u(x,t) = 0$ for all $x\in\mathbb{R}^3$ and, say, for $-5/6 < t < 0$. Since u is divergence free, it is a harmonic function in $\mathbb{R}^3$ depending on $t\in]-5/6,0[$ as a parameter. Therefore, for any $a > \sqrt{5/6}$ and for any $x_0\in\mathbb{R}^3$, by the mean value theorem for harmonic functions, we have

$$\begin{aligned}
&\sup_{-5/6<t<0}|u(x_0,t)|^2 \le\\
&c\sup_{-5/6<t<0}\frac{1}{a^3}\int\limits_{B(x_0,a)}|u(x,t)|^2dx \le\\
&c\sup_{-5/6<t<0}\frac{1}{a^3}\int\limits_{B(|x_0|+a)}|u(x,t)|^2dx \le\\
&c\frac{a+|x_0|}{a^3}A(u,a+|x_0|).
\end{aligned}$$

Thanks to boundedness of scaled energy quantities stated in Proposition 3.1,

the right hand side of the latter inequality tends to zero as a goes to infinity. By arbitrariness of x_0, we conclude that $u(x,t) = 0$ for all $x \in \mathbb{R}^3$ and for $-5/6 < t < 0$, which contradicts (3.1). Hence, the origin $z = 0$ cannot be a singular point of v.

Let us go back to the marginal case of Ladyzhenskaya-Prodi-Serrin condition, the so-called $L_{3,\infty}$-case, and show that it is completely embedded into the above scheme. So, we assume that functions v and q are a suitable weak solution to the Navier-Stokes equations in Q and satisfy the additional condition

$$\|v\|_{3,\infty,Q} < +\infty. \tag{3.8}$$

With the help of Proposition 1.3, it is not so difficult to show that $g' < +\infty$. So, for v, all the assumptions of Proposition 3.1 hold and thus our blowup procedure produces a local energy ancient solution u with the properties listed in that proposition. Exploited the ε-regularity theory once more, we can show further that $v(\cdot,0) \in L_3(B(2/3))$, which in turn implies (3.2). Now, in order to prove regularity of the velocity v at the point $z = 0$, it is sufficient to verify the validity of (3.7). Indeed, by scale-invariance,

$$\|u\|_{3,\infty,\mathbb{R}^3\times]-\infty,0[} < +\infty.$$

Applying Proposition 3.1 once again and taking into account properties of harmonic functions, one can conclude that

$$\|p\|_{\frac{3}{2},\infty,\mathbb{R}^3\times]-\infty,0[} < +\infty.$$

Combining the latter estimates, we show that for any $T > 0$

$$\int\limits_{-T}^{0}\int\limits_{\mathbb{R}^3} \left(|u|^3 + |p|^{\frac{3}{2}}\right) dxdt < +\infty. \tag{3.9}$$

Our further arguments rely upon the ε-regularity theory. Indeed, letting, say, $T = 4$, one can find $R > 4$ so that

$$\int\limits_{-4}^{0}\int\limits_{\mathbb{R}^3\setminus B(R/2)} \left(|u|^3 + |p|^{\frac{3}{2}}\right) dxdt < \varepsilon.$$

The rest of the proof of (3.7) is easy.

4. How Does L_3-norm Approach Potential Blowup?

Let v be a weak Leray-Hopf solution to the classical Cauchy problem. Assume that it has a finite blowup time T. As it has been already shown by J. Leray,

for any $3 < s \le +\infty$, there is a positive constant c_s such that

$$\|v(\cdot,t)\|_{s,\mathbb{R}^3} \ge \frac{c_s}{(T-t)^{\frac{s-3}{2s}}}$$

for $t < T$.

However, in the limit case $s = 3$, according to what has been discussed in the previous section, we just have

$$\limsup_{t\to T-0} \|v(\cdot,t)\|_{3,\mathbb{R}^3} = +\infty.$$

It would be natural to ask whether

$$\lim_{t\to T-0} \|v(\cdot,t)\|_{3,\mathbb{R}^3} = +\infty \tag{4.1}$$

is true or not? An answer to this question is still unknown. In this section, we shall seek either some additional conditions providing the positive answer or a weaker version of (4.1). Regarding to the first goal, we could formulate the following statement.

Theorem 4.1. *Let v be a weak Leray-Hopf solution to the Cauchy problem (1.1) and (1.2) and let T be a finite time blowup. Assume that, for some $3 < s \le +\infty$, there exists a positive constant C_s such that*

$$\|v(\cdot,t)\|_{s,\mathbb{R}^3} \le \frac{C_s}{(T-t)^{\frac{s-3}{2s}}}$$

for $t < T$. Then (4.1) is true.

Let us outline the proof of Theorem 4.1 following [30]. Without loss of generality, we always may assume $s = +\infty$, which means that we restrict ourselves to the case of Type I blowups. Indeed, this is a simple consequence of the ε-regularity theory. So, after application of Proposition 1.3, we can conclude that $g' < +\infty$.

It is known that, at the blowup time, all singular points of any weak Leray-Hopf solution v belong to a bounded ball whose radius in a way depends upon v. So, just by shift in space-time, we may assume that the origin $z = 0$ is a singular point of v. Suppose further that (4.1) is wrong. Then, there exists an increasing sequence $\{t_k\}_{k=1}^{\infty}$ tending to zero such that

$$\sup_k \|v(\cdot,t_k)\|_{3,\mathbb{R}^3} = M < +\infty. \tag{4.2}$$

Making use of Proposition 3.1, we may construct a local energy ancient solution u. The sequence λ_k will be specified later. By partial regularity theory and by

(4.2), one can assert that $\|v(\cdot,0)\|_{3,\mathbb{R}^3} < +\infty$, which in turn implies identity (3.3) for u. In addition, by scale-invariance, our local energy ancient solution satisfies the estimate

$$\|u(\cdot,s)\|_{\infty,\mathbb{R}^3} \le \frac{C_\infty}{\sqrt{-s}} \tag{4.3}$$

for any $-\infty < s < 0$.

Now, we need to provide the validity of decay estimate (3.7). To this end, we choose λ_k in a special way

$$\lambda_k = \frac{\sqrt{-t_k}}{2}.$$

In the rest of this section, we shall show why such a choice of λ_k gives (3.7). The first argument in support of it is as follows: u obeys the important global property

$$u(\cdot,-4) \in L_3(\mathbb{R}^3). \tag{4.4}$$

Our next step is to construct a solution to the Cauchy problem for the Navier-Stokes equations with the velocity field $u(\cdot,-4)$ as the initial datum, i.e., to solve the following initial value problem

$$\begin{aligned} \partial_t w(x,t) + w(x,t)\cdot\nabla w(x,t) - \Delta w(x,t) + \nabla r &= 0, \\ \operatorname{div} w(x,t) &= 0 \end{aligned} \tag{4.5}$$

for $x \in \mathbb{R}^3$ and $-4 < t < 1$,

$$w(x,-4) = u(x,-4)$$

for $x \in \mathbb{R}^3$. With such initial data, one can find a mild solution (in Kato's sense) but in general it is local in time and thus does not necessary cover the whole time interval $]-4,1[$. On the other hand, we cannot ensure the existence of a weak Leray-Hopf solution since $u(\cdot,-4)$ is not necessary in $L_2(\mathbb{R}^3)$.

The way out is to use an interesting conception of local energy solutions introduced by P. G. Lemarié-Riesset in [21]. This is an important generalization of the notion of weak Leray-Hopf solutions to the Cauchy problem. The phase space for local energy solutions is defined with the help of a particular Morrey space

$$L_{2,unif} = \left\{ \|u\|_{L_{2,unif}} = \sup_{x\in\mathbb{R}^3} \|u\|_{2,B(x,1)} < +\infty \right\}.$$

To proceed with our definitions, let us find the completion of $C_0^\infty(\mathbb{R}^3)$ in $L_{2,unif}$ and denote it by E_2. It is not so difficult to check that the space $L_3(\mathbb{R}^3)$ is embedded into the space E_2. Finally, the phase (energy) space $\overset{\circ}{E}_2$ consists of

all divergence free vector fields belonging to E_2. Having such an energy space in hands, one can give a complete definition of local energy solutions. In our version, we follow paper [15]. It is a little modification of the original definition adopted in the monograph [21].

Definition 4.2. *Functions $w \in L_\infty(-4,1;\overset{\circ}{E}_2)$ and $r \in L_{\frac{3}{2}}(-4,1;L_{\frac{3}{2},loc}(\mathbb{R}^3))$ are said to be a local energy weak Leray-Hopf solution or simply local energy solution to the Cauchy problem (4.5) if the following conditions hold:*

$$\sup_{x_0\in\mathbb{R}^3}\int\limits_{-4}^{1}\int\limits_{B(x_0,1)}|\nabla w|^2dz<+\infty,$$

w and r meet (4.5) in the sense of distributions;

the function $t\mapsto\int\limits_{\mathbb{R}^3}w(x,t)\cdot\widetilde{w}(x)\,dx$ is continuous on $[-4,1]$ for any compactly supported function $\widetilde{w}\in L_2(\mathbb{R}^3)$;

for any compact K,

$$\|w(\cdot,t)-u(\cdot,-4)\|_{L_2(K)}\to 0$$

as $t\to -4+0$;

for a.a. $t\in]-4,1[$, the local energy inequality

$$\int\limits_{\mathbb{R}^3}\varphi|w(x,t)|^2\,dx+2\int\limits_{-4}^{t}\int\limits_{\mathbb{R}^3}\varphi|\nabla w|^2\,dxdt'\le\int\limits_{-4}^{t}\int\limits_{\mathbb{R}^3}\Big(|w|^2(\partial_t\varphi+\Delta\varphi)+w\cdot\nabla\varphi(|w|^2+2r)\Big)\,dxdt'$$

is valid for all nonnegative functions $\varphi\in C_0^\infty(\mathbb{R}^3\times]-4,2[)$;

for each point $x_0\in\mathbb{R}^3$, there exists a function $c_{x_0}\in L_{\frac{3}{2}}(-4,1)$ such that

$$r_{x_0}(x,t)\equiv r(x,t)-c_{x_0}(t)=r^1_{x_0}(x,t)+r^2_{x_0}(x,t),$$

for $(x,t)\in B(x_0,3/2)\times]-4,1[$, where

$$r^1_{x_0}(x,t)=-\frac{1}{3}|w(x,t)|^2+\frac{1}{4\pi}\int\limits_{B(x_0,2)}K(x-y):w(y,t)\otimes w(y,t)\,dy,$$

$$r^2_{x_0}(x,t)=\frac{1}{4\pi}\int\limits_{\mathbb{R}^3\setminus B(x_0,2)}(K(x-y)-K(x_0-y)):w(y,t)\otimes w(y,t)\,dy,$$

and $K(x)=\nabla^2(1/|x|)$.

As in the case of weak Leray-Hopf solutions, uniqueness of local energy solutions to the Cauchy problem (4.5) is an open problem. However, bound (4.3) makes it possible to show that our ancient solution u and any local energy solution w, defined by the Cauchy problem (4.5) with the help of u, coincide in the time interval $]-4,0[$, see a proof, for example, in [30]. On the other hand, we would like to remind that local energy solutions satisfy the local energy inequality and thus the ε-regularity theory is applicable to them. Making use of the ε-regularity theory, together with a certain decay of some integral norms over unit balls with respect to their centers, see [21] or [15] for exact statements and for a proof, one can show required point-wise estimate (3.7) for w and thus for u. The rest of the proof is more or less the same as in the local $L_{3,\infty}$-case, which means that $z=0$ is actually not a singular point.

Now we are going to discuss a weaker version of (4.1).

Theorem 4.3. *Let v be a weak Leray-Hopf solution to the Cauchy problem (1.1) and (1.2) and let T be a finite time blowup. Then*

$$\lim_{t\to T-0}\frac{1}{T-t}\int\limits_t^T \|v(\cdot,\tau)\|^3_{3,\mathbb{R}^3}d\tau=+\infty.$$

Apparently, Theorem 4.3 can be easily deduced from its local version, see [29], which reads:

Proposition 4.4. *Let v and q be a suitable weak solution to the Navier-Stokes equations in Q. Assume, in addition, that*

$$\liminf_{t\to -0}\frac{1}{-t}\int\limits_t^0 \|v(\cdot,\tau)\|^3_{3,B}d\tau<+\infty. \tag{4.6}$$

Then $z=0$ is a regular point of v.

Let us outline a proof of Proposition 4.4. There are two simple consequences of (4.6):

$$M=\sup_k\frac{1}{-t_k}\int\limits_{t_k}^0\int\limits_B |v(x,\tau)|^3dxd\tau<+\infty$$

for some increasing sequence $\{t_k\}$ tending to zero and

$$v(\cdot,0)\in L_3(Q(5/6)). \tag{4.7}$$

As to estimates for the pressure field, one can split it into two parts so that

$q = q^1 + q^2$ where q^1 has a "good" bound

$$\frac{1}{-t_k}\int\limits_{t_k}^{0}\int\limits_{B}|q^1(x,\tau)|^{\frac{3}{2}}dxd\tau \le cM$$

with a universal constant c while the second counter-part is a harmonic function satisfying the estimate

$$\sup_{x\in B(2/3)}|q^2(x,t)|^{\frac{3}{2}} \le c\left(\int\limits_{B}|q(x,t)|^{\frac{3}{2}}dx + \int\limits_{B}|v(x,t)|^3dx\right).$$

We cannot apply Proposition 3.1 directly since we do not know whether g' is bounded or not. So, we have to prove the existence of a non-trivial blowup solution by hands.

We can choose $\lambda_k = \sqrt{-t_k/10}$ and then, just by scaling, pick up subsequences, denoted by the same symbols, such that, for any positive number a,

$$u^{(k)} \rightharpoonup u$$

in $L_3(B(a)\times]-10,0[)$,

$$p^{1(k)} \rightharpoonup p$$

in $L_{\frac{3}{2}}(B(a)\times]-10,0[)$, and

$$p^{2(k)} \to 0$$

in $L_{\frac{3}{2}}(B(a)\times]-10,0[)$. Here, $p^{i(k)}(y,s) = \lambda_k^2 q^i(\lambda_k y, \lambda_k^2 s)$, $i = 1,2$. Then by the linear theory and by the usual compactness arguments, one can show that, for any $a > 0$,

$$u^{(k)} \to u$$

in $L_3(B(a)\times]-10,0[)\cap C([-10,0];L_{\frac{9}{8}}(B(a)))$. Moreover, function u and p are a suitable weak solution to the Navier-Stokes equations in $\mathbb{R}^3\times]-10,0[$ that has the properties $u \in L_3(\mathbb{R}^3\times]-10,0[)$ and $p \in L_{\frac{3}{2}}(\mathbb{R}^3\times]-10,0[)$. This solution is non-trivial since it satisfies (3.1). The further arguments are similar to those which have been used in Section 3. Finally, (4.7) implies (3.3). Repeating the end of the proof in $L_{3,\infty}$-case, one can state that, in fact, u is identically equal to zero, say, on $\mathbb{R}^3\times]-1,0[$, which contradicts (3.1). So, the origin is a regular point of v.

References

[1] L. Caffarelli, R.-V. Kohn, L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier-Stokes equations*, Comm. Pure Appl. Math., **XXXV** (1982), 771–831.

[2] M. Cannone, *Harmonic analysis tools for solving the incompressible Navier-Stokes equations*, Handbook of Mathematical Fluid Dynamics, Edited by Friedlander, D. Serre, **3** (2004), 161–244.

[3] C.-C. Chen, R. M. Strain, T.-P. Tsai, H.-T. Yau, *Lower bound on the blow-up rate of the axisymmetric Navier-Stokes equations*, Int. Math. Res. Not., (2008) Vol. 2008 : article ID rnn016, 31 pages, doi:10.1093/imrn/rnn016.

[4] C.-C. Chen, R. M. Strain, T.-P. Tsai, H.-T. Yau, *Lower bounds on the blow-up rate of the axisymmetric Navier-Stokes equations II*, Communications in PDE's, **34**(2009), 203–232.

[5] Z.-M. Chen, W. G. Price, *Blow-up rate estimates for weak solutions of the Navier–Stokes equations*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **457** (2001), 2625–2642.

[6] H. L. Choe, J. L. Lewis, *On the singular set in the Navier-Stokes equations*, J. Functional Anal., **175** (2000), 348–369.

[7] L. Escauriaza, c. E. Kenig, G. Ponce, L. Vega, *Decay at infinity of caloric functions within characteristic hyperplanes*, Math. Res. Lett., **13** (2006), 441-453.

[8] L. Escauriaza, G. Seregin, V. Šverák, *$L_{3,\infty}$-solutions to the Navier-Stokes equations and backward uniqueness*, Russian Mathematical Surveys, **58** (2003), 211–250.

[9] L. Escauriaza, G. Seregin, V. Šverák, *On backward uniqueness for parabolic equations*, Arch. Rational Mech. Anal., **169** (2003), 147–157.

[10] C. Fefferman, http://www.claymath.org/millennium/Navier-Stokes equations.

[11] Y. Giga, *Solutions for semilinear parabolic equations in L^p and regularity of weak solutions of the Navier-Stokes equations*, J. of Diff. Equations, **62** (1986), 186–212.

[12] S. Gustafson, K. Kang, T.-P. Tsai, *Interior regularity criteria for suitable weak solutions of the Navier-Stokes equations*, Commun. Math. Phys., **273** (2007), 161–176.

[13] E. Hopf, *Über die Anfangswertaufgabe für die hydrodynamischen Grundgleichungen*, Math. Nachrichten, **4** (1950-51), 213–231.

[14] T. Kato, *Strong L_p-solutions of the NavierStokes equation in R^m, with applications to weak solutions*, Math. Z., **187** (1984), 471-480.

[15] N. Kikuchi, G. Seregin, *Weak solutions to the Cauchy problem for the Navier-Stokes equations satisfying the local energy inequality.* AMS translations, Series 2, **220**, 141–164.

[16] H. Kim, H. Kozono, *Interior regularity criteria in weak spaces for the Navier-Stokes equations*, manuscripta math., **115** (2004), 85-100.

[17] G. Koch, N. Nadirashvili, G. Seregin, V. Šverák, *Liouville theorems for the Navier-Stokes equations and applications*, Acta Mathematica, **203** (2009), 83–105.

[18] H. Koch, D. Tataru, *Well-posedness for the NavierStokes equations*, Adv. Math., **157** (2001), 22-35.

[19] O. A. Ladyzhenskaya, *Mathematical problems of the dynamics of viscous incompressible fluids*, 2nd edition, Nauka, Moscow 1970.

[20] O. A. Ladyzhenskaya, *On unique solvability of the three-dimensional Cauchy problem for the Navier-Stokes equations under the axial symmetry*, Zap. Nauchn. Sem. LOMI **7** (1968), 155–177.

[21] P. G. Lemarié-Riesset, *Recent developemnets in the Navier-Stokes problem*, Chapman&Hall/CRC reseacrh notes in mathematics series, 431.

[22] J. Leray, *Sur le mouvement d'un liquide visqueux emplissant l'espace*, Acta Math. **63** (1934), 193–248.

[23] F.-H. Lin, *A new proof of the Caffarelly-Kohn-Nirenberg theorem*, Comm. Pure Appl. Math., **51** (1998), 241–257.

[24] J. Necas, M. Ruzicka, V. Šverák, On Leray's self-similar solutions of the Navier-Stokes equations. Acta Math. **176** (1996), 283–294.

[25] V. Scheffer, *Partial regularity of solutions to the Navier-Stokes equations*, Pacific J. Math., **66** (1976), 535–552.

[26] V. Scheffer, *Hausdorff measure and the Navier-Stokes equations*, Commun. Math. Phys., **55** (1977), 97–112.

[27] G. Seregin, *Local regularity theory of the Navier-Stokes equations*, Handbook of Mathematical Fluid Mechanics, Edited by Friedlander, D. Serre, **4** (2007), 159–200.

[28] G. Seregin, *Local regularity for suitable weak solutions of the NavierStokes equations*, Russian Math. Surveys **62** (2007), 595–614.

[29] G. A. Seregin, *Navier-Stokes equations: almost $L_{3,\infty}$-cases*, Journal of mathematical fluid mechanics, **9** (2007), 34–43.

[30] G. Seregin, *A note on necessary conditions for blow-up of energy solutions to the Navier-Stokes equations*, arXiv:0909.3897.

[31] G. Seregin, V. Šverák, *On Type I singularities of the local axi-symmetric solutions of the Navier-Stokes equations*, Communications in PDE's, **34** (2009), 171–201.

[32] G. Seregin, W. Zajaczkowski, *A sufficient condition of local regularity for the Navier-Stokes equations*, Zapiski Nauchn. Seminar, POMI, **336** (2006), 46–54.

[33] G. Seregin, W. Zajaczkowski, *A sufficient condition of regularity for axially symmetric solutions to the Navier-Stokes equations*, SIMA J. Math. Anal., **39**(2007), 669–685.

[34] J. Serrin, *On the interior regularity of weak solutions of the Navier-Stokes equations*, Arch. Ration. Mech. Anal., **9** (1962), 187–195.

[35] M. Struwe *On partial regularity results for the NavierStokes equations*, Comm. Pure Appl. Math., **41** (1988), 437-458.

[36] S. Takahashi, *On interior regularity criteria for weak solutions of the Navier-Stokes equations*, Manuscripta Math., **69** (1990), 237–254.

[37] M. R. Ukhovskij, V. L. Yudovich, *Axially symmetric motions of ideal and viscous fluids filling all space*, Prikl. Mat. Mech. **32** (1968), 59–69.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Weakly Nonlinear Wave Equations with Random Initial Data

Herbert Spohn*

Abstract

We discuss the derivation of the kinetic equation for the weakly nonlinear Schrödinger equation on the lattice $\mathbb{Z}^d$ and state a theorem, which establishes that the equilibrium time covariance is damped because of the nonlinearity. A more general space-time central limit theorem is discussed.

Mathematics Subject Classification (2010). Primary 82C05; Secondary 35Q55.

Keywords. Kinetic theory of wave equations

1. Introduction

The statistical mechanics of time-dependent phenomena deals with systems of many degrees of freedom and random initial data. In the classical context the evolution equations are deterministic and of Hamiltonian type. The initial data are distributed according to some probability measure on phase space. The latter assumption has stirred an ongoing debate centered around the problem of the origin of this randomness. The proposed answers have a wide spectrum, see [16] and [14] Chapter 5 and 7. One approach tries to relate the assumed random initial data *now* to some properties of the system, possibly enlarged through interactions with an environment, at some earlier time. The other extreme is to regard the random initial data merely as a mathematical device, because of the impossibility to follow individual trajectories. Of course, in such a scheme one takes the responsibility to demonstrate that the predicted properties do not depend too sensitively on the particular choice of the initial measure. For example, one could argue that the chaoticity of the dynamics washes out the information on finer details of the initial conditions. In the present lecture

*I report on joint work with Jani Lukkarinen, Dep. Mathematics, Helsinki University.

Zentrum Mathematik and Physik Department, TU München, D-85747 Garching, Germany. E-mail: spohn@ma.tum.de.

we will not touch upon such issues and pursue the pragmatic attitude that reasonable initial probability measures will be delineated through the physical context. Notwithstanding, initial randomness will be mandatory. A pointwise, in the initial conditions, result is highly unlikely. At best one can hope that the behavior claimed below holds for "typical" initial conditions.

More specifically, the topic of the lecture are weakly nonlinear wave equations with random initial data, see [21, 19] for the nonlinear case and [17, 2] for the linear case with random coefficients. Many of our considerations are valid for a general class of wave equations. But only the nonlinear Schrödinger equation has been studied in some mathematical detail and we will focus exclusively on this case.

The wave field, $\psi(x,t)$, $x \in \mathbb{R}^d$, $t \in \mathbb{R}$, is complex valued and governed by the evolution equation

$$\mathrm{i}\frac{\partial}{\partial t}\psi(x,t) = -\tfrac{1}{2}\Delta\psi(x,t) + \lambda|\psi(x,t)|^2\psi(x,t)\,, \tag{1.1}$$

which is to be solved as Cauchy problem. We study only the defocusing case $\lambda \geq 0$, and eventually $\lambda \ll 1$. Instead of $\mathbb{R}^d$ one could also consider some bounded domain $\Lambda \subset \mathbb{R}^d$ with smooth boundary and would have to prescribe suitable boundary values at $\partial\Lambda$.

(1.1) is of Hamiltonian form. The Hamiltonian function is

$$H(\psi) = \tfrac{1}{2}\int_{\mathbb{R}^d}\big(|\nabla\psi(x)|^2 + \lambda|\psi(x)|^4\big)\mathrm{d}x\,. \tag{1.2}$$

Defining the conjugate canonical fields by

$$q(x) = \frac{1}{\sqrt{2}}\Re\psi(x)\,, \quad p(x) = \frac{1}{\sqrt{2}}\Im\psi(x)\,, \tag{1.3}$$

Hamilton's equation of motion for (1.2) yield the evolution equation (1.1). The nonlinear Schrödinger equation is singled out from other wave equations because q and p appear symmetrically in H. This symmetry simplifies the estimates considerably.

The solution to (1.1) conserves the $L^2(\mathbb{R}^d)$-norm (the number) and the energy

$$\frac{d}{dt}\|\psi(t)\|^2 = 0\,, \quad \frac{d}{dt}H(\psi(t)) = 0\,. \tag{1.4}$$

Therefore a natural phase space for the nonlinear Schrödinger equation (1.1) is $\Gamma = \{\psi \in L^2(\mathbb{R}^d)|H(\psi) < \infty\}$. In this space (1.1) has solutions global in time.

The initial, $t = 0$, data are distributed according to a probability measure μ on Γ. The prototypical example is a gauge invariant Gaussian measure μ, where the former condition means that for every $\vartheta \in [0, 2\pi]$ it holds

$$\psi(x) = \mathrm{e}^{\mathrm{i}\vartheta}\psi(x) \tag{1.5}$$

in probability. The Gaussian measure is then uniquely characterized by the expectations

$$\mathbb{E}_\mu(\psi(x)) = 0\,, \quad \mathbb{E}_\mu(\psi(x)\psi(y)) = 0\,, \tag{1.6}$$

and the nontrivial part of the covariance

$$\mathbb{E}_\mu(\langle f,\psi\rangle\langle\psi,g\rangle) = \langle f, Cg\rangle\,. \tag{1.7}$$

Here $\langle\cdot,\cdot\rangle$ is the scalar product of $L^2(\mathbb{R}^d)$, $f,g \in L^2(\mathbb{R}^d)$, and $C \geq 0$ with $\mathrm{tr}\, C < \infty$, $\mathrm{tr}[(-\Delta)C] < \infty$, $\int C(x,x)^2 \mathrm{d}x < \infty$. The latter conditions ensure that μ is supported by Γ.

Generally speaking, one would like to understand the qualitative behavior of solutions, in particular their long time limit. With current techniques such a goal is too ambitious. However, there is some progress in understanding the solution behavior for small λ.

2. Finite *vs.* Infinite Volume, Continuum *vs.* Lattice, Equilibrium Measures

We formulated the Cauchy problem (1.1) in $\mathbb{R}^d$ and with finite number and energy. Thus one expects the solution to spread out as $t \to \infty$. A more interesting long time behavior should result in case the system is confined to a box. This brings in some dependence on boundary conditions. Another, from the perspective of statistical mechanics very natural option would be to discard the condition $\mathrm{tr}\, C < \infty$ and to choose an initial measure μ, which is Gaussian, gauge invariant, and translation invariant, thus determined by

$$\mathbb{E}_\mu(\psi(x)) = 0\,, \quad \mathbb{E}_\mu(\psi(x)\psi(y)) = 0\,, \tag{2.1}$$

$$\mathbb{E}_\mu(\psi(x)^*\psi(y)) = (2\pi)^{-d}\int \mathrm{e}^{\mathrm{i}k\cdot(x-y)} W(k)\mathrm{d}k \tag{2.2}$$

with W smooth and of rapid decay. For such a measure, the typical ψ field has a logarithmic increase at infinity. Thus the issue of existence of global solution arises, which to our knowledge has not yet been studied. In fact to deal with this problem, it might be simpler to mollify the $|\psi|^4$ nonlinearity by

$$V(\psi) = \tfrac{1}{2}\lambda \int |\psi(x)|^2 V(x-y)|\psi(y)|^2 \mathrm{d}x\mathrm{d}y\,, \tag{2.3}$$

$V : \mathbb{R}^d \to \mathbb{R}$, $V(x) = V(-x)$, V continuous, V of finite range, and $\widehat{V} \geq 0$.

Out of the set of all probability measures, physically singled out are those describing thermal equilibrium. Of course, such a measure is stationary in time. But statistical correlations in equilibrium depend on space *and* time and one

may ask about their properties in case of weak nonlinearity. The equilibrium measure in some box $\Lambda \subset \mathbb{R}^d$ is formally given by the Gibbs measure

$$\frac{1}{\mathcal{N}} \prod_{x\in\Lambda} \left(\mathrm{d}\Re\psi(x)\mathrm{d}\Im\psi(x)\right) \exp[-\beta H(\psi) + \beta\lambda_0 N(\psi)] \tag{2.4}$$

with $\lambda_0 < 0$. To construct such a measure is one central problem of Euclidean Quantum Field Theory [18]. Even with (2.3) instead of $|\psi|^4$ the construction is difficult. One also would have to show that the dynamics exists for a set of initial fields which has full probability in equilibrium.

On general grounds, it is a wise strategy not to mix essentially disjoint problems. Following this guide line, in our case the natural choice is to discretize (1.1) and to thereby remove all ultraviolet difficulties. In particular, the dynamics is well defined even for data with a slow increase at infinity and the equilibrium measure (2.4) exists at infinite volume without any further limiting procedure. There is a small price to pay, however: The linear part of the wave equation requires more attention because of lattice effects. Still we replace $\mathbb{R}^d$ by $\mathbb{Z}^d$, to be kept as a standing assumption. The wave field is a function on $\mathbb{Z}^d$, $\psi : \mathbb{Z}^d \to \mathbb{C}$, the energy is given by

$$H(\psi) = \sum_{x,y\in\mathbb{Z}^d} \alpha(x-y)\psi(x)^*\psi(y) + \tfrac{1}{2}\lambda \sum_{x,y\in\mathbb{Z}^d} |\psi(x)|^2 V(x-y)|\psi(y)|^2\,, \tag{2.5}$$

and the number by

$$N(\psi) = \sum_{x\in\mathbb{Z}^d} |\psi(x)|^2\,. \tag{2.6}$$

Here the real couplings α satisfy $\alpha(x) = \alpha(-x)$ and $\alpha(x) = 0$ for $|x| > r$, the range of α. For the discrete Laplacian one would set $\alpha(x) = -1/2d$ for $|x| = 1$ and $\alpha(x) = 0$ otherwise. But it is convenient to allow for a general α. The equations of motion now read

$$\mathrm{i}\frac{d}{dt}\psi_t(x) = \sum_{y\in\mathbb{Z}^d} \alpha(x-y)\psi_t(y) + \lambda \sum_{y\in\mathbb{Z}^d} |\psi_t(y)|^2 V(y-x)\psi_t(x)\,. \tag{2.7}$$

It is instructive to rewrite the dynamics in Fourier space. For this purpose let us denote the Fourier transform of $f : \mathbb{Z}^d \to \mathbb{C}$ by

$$\widehat{f}(k) = \sum_{x\in\mathbb{Z}^d} f(x)\mathrm{e}^{-\mathrm{i}2\pi k\cdot x}\,, \tag{2.8}$$

$k \in \mathbb{R}$, and the inverse Fourier transform by

$$\widetilde{g}(x) = \int_{\mathbb{T}^d} g(k)\mathrm{e}^{\mathrm{i}2\pi k\cdot x}\mathrm{d}k \tag{2.9}$$

with $\mathbb{T}^d = [0,1]^d$, as one parametrization of the d-dimensional torus. $\langle\cdot,\cdot\rangle$ is now the inner product for $\ell_2(\mathbb{Z}^d)$, resp. for $L^2(\mathbb{T}^d)$, depending on the context. The dispersion relation is defined by

$$\omega(k) = \widehat{\alpha}(k) \tag{2.10}$$

and has the properties

(1) $\omega : \mathbb{T}^d \to \mathbb{R}$ and its periodic extension are real analytic functions.
(2) $\omega(k) = \omega(-k)$.

Written in Fourier space the energy is

$$H(\psi) = \int_{\mathbb{T}^d} \omega(k)|\widehat{\psi}(k)|^2 \mathrm{d}k + \tfrac{1}{2}\lambda \int_{(\mathbb{T}^d)^4} \delta(k_1+k_2-k_3-k_4)\widehat{V}(k_2-k_3) \times \widehat{\psi}(k_1)^* \widehat{\psi}(k_2)^* \widehat{\psi}(k_3)\widehat{\psi}(k_4)\mathrm{d}k_1\mathrm{d}k_2\mathrm{d}k_3\mathrm{d}k_4 \tag{2.11}$$

and the equations of motion read

$$\frac{d}{dt}\widehat{\psi}_t(k_1) = -\mathrm{i}\omega(k_1)\widehat{\psi}_t(k_1) - \mathrm{i}\lambda \int \delta(k_1+k_2-k_3-k_4)\widehat{V}(k_2-k_3) \times \widehat{\psi}_t(k_2)^* \widehat{\psi}_t(k_3)\widehat{\psi}_t(k_4)\mathrm{d}k_2\mathrm{d}k_3\mathrm{d}k_4\,. \tag{2.12}$$

In the discrete setting the existence and uniqueness of global solutions is ensured, as first proved by Lanford, Lebowitz, and Lieb [12] in a comparable context. Sharper propagation estimates and a controlled infinite volume limit is achieved by Buttà *et al.* [5], on which the theorem below is based. One has to assume the stability condition $\widehat{V}(k) \geq 0$. To state the theorem we introduce the local energy in the cube $\Lambda_{\nu,\ell}$ of center ν and side-length $2\ell+1$,

$$H_{\nu,\ell}(\psi) = \sum_{x,y\in\Lambda_{\nu,\ell}} \psi(x)^*\alpha(x-y)\psi(y) + \sum_{x\in\Lambda_{\nu,\ell}} (\lambda_0|\psi(x)|^2 + 1) + \lambda \sum_{x,y\in\Lambda_{\nu,\ell}} |\psi(x)|^2 V(x-y)|\psi(y)|^2 \tag{2.13}$$

and the "average" maximal energy

$$Q(\psi) = \sup_{\nu\in\mathbb{Z}^d} \sup_{\ell \geq \left(\log(\mathrm{e}+|\nu|)\right)^{1/d}} (2\ell+1)^{-d} H_{\nu,\ell}(\psi)\,. \tag{2.14}$$

Here λ_0 is determined such that $\widehat{\alpha}(k) + \lambda_0 > 0$. Let

$$\aleph = \{\psi | Q(\psi) < \infty\}\,. \tag{2.15}$$

Theorem 2.1. *Let $\widehat{V}(k) \geq 0$ and $\lambda \geq 0$. Then there exists a one-parameter group of transformations $\Phi_t : \aleph \to \aleph$, $t \in \mathbb{R}$, such that $t \mapsto \Phi_t(\psi)$ is the unique global solution to Eq. (2.7) with initial conditions $\Phi_0(\psi) = \psi$.*

The set $\aleph$ is sufficiently large to support all initial measures of physical interest, in particular the Gaussian measure μ with covariance (2.2).

3. Kinetic Limit

For the nonlinear Schrödinger equation (2.7) we fix λ, but $\lambda \ll 1$. Then for times at least up to order λ^{-1} the nonlinearity can be ignored. Thus for a while we set $\lambda = 0$ and will come back to the issue of how to properly incorporate the nonlinearity. Let us denote by $U(t)$ the flow corresponding to the solution of (2.7) with $\lambda = 0$. Let us impose good spatial mixing for the initial measure μ. Then because of finite speed of propagation the random process $t \to (U(t)\psi)(x)$ depends for different times on essentially disjoint sets of initial ψ's. Hence the good mixing in space translates to a good mixing in time, which implies that the suitably averaged process $U(t)\psi(x)$ is approximately Gaussian. This common lore has been worked out in a few exemplary cases. One is a noninteracting Fermi liquid on a lattice [11]. Another class are discretized wave equations, *alias* harmonic crystals [10, 6].

Most easily stated is the spatially homogeneous case, see [11]. Let μ be a translation invariant probability measure on $\mathbb{C}^{\mathbb{Z}^d}$ with mean zero, $\mathbb{E}_\mu(\psi(x)) = 0$. The spatial mixing is formalized by the property of ℓ_1-clustering which states that the fully truncated correlation functions satisfy the ℓ_1-bound

$$\sum_{(x_1,\dots,x_n)\in(\mathbb{Z}^d)^n} \delta_{x_1,0} \Big| \mathbb{E}_\mu \left(\prod_{j=1}^{n} \psi(x_j, \sigma_j) \right)_{\mathrm{T}} \Big| < \infty \tag{3.1}$$

for every $n \geq 2$. Here the subscript T means full truncation and we use the convention $\psi(x,1) = \psi(x)$, $\psi(x,-1) = \psi(x)^*$, hence $\sigma_j = \pm 1$. Let

$$\mathbb{E}_\mu\big(\psi(x)^*\psi(y)\big) = C(x-y) \tag{3.2}$$

with $\widehat{C}$ bounded and continuous and let μ_{G} be the translation and gauge invariant Gaussian measure with the covariance (3.2). Then, in the sense of convergence of moments,

$$\lim_{t\to\infty} \mu \circ U(-t) = \mu_{\mathrm{G}}\,. \tag{3.3}$$

The proof requires the decay of certain ℓ_p bounds on the spatial propagator

$$p_t(x) = \int_{\mathbb{T}^d} \mathrm{e}^{\mathrm{i}2\pi k\cdot x} \mathrm{e}^{-\mathrm{i}\omega(k)t} \mathrm{d}k\,. \tag{3.4}$$

Stronger general results, assuming probabilistic mixing conditions for μ, are discussed in [6].

In the kinetic context one would like to treat also some spatial variation. This can be most easily formulated by introducing the Wigner function W_ψ, which filters out the slowly varying pieces of products as $\psi^*\psi$. To each ψ we associate

$$W_\psi^\varepsilon(y,k) = (\varepsilon/2)^d \int_{(2\mathbb{T}/\varepsilon)^d} \mathrm{e}^{\mathrm{i}2\pi y\cdot\eta} \widehat{\psi}(k - \tfrac{1}{2}\varepsilon\eta)^* \widehat{\psi}(k + \tfrac{1}{2}\varepsilon\eta) \mathrm{d}\eta \tag{3.5}$$

with $k \in \mathbb{T}^d$, $y \in (\varepsilon\mathbb{Z}/2)^d$, and ε a dimensionsless scale parameter, $\varepsilon > 0$ and $\varepsilon \ll 1$. We prescribe a sequence μ_ε of probability measures and assume that the average Wigner function has a limit as $\varepsilon \to 0$,

$$\lim_{\varepsilon\to 0} \mathbb{E}_{\mu_\varepsilon}\big(W^\varepsilon_\psi(\lfloor r\rfloor_\varepsilon, k)\big) = W(r,k) \tag{3.6}$$

with some bounded and smooth $W : \mathbb{R}^d \times \mathbb{T}^d \to \mathbb{R}$. Here $\lfloor\cdot\rfloor_\varepsilon$ denotes the integer part modulo $(\varepsilon\mathbb{Z}/2)^d$. If the limit in (3.6) exists, then the sequence of measures μ_ε have a slow spatial variation of order ε^{-1} on the scale of the lattice $\mathbb{Z}^d$.

Slow variation is a macroscopic property and therefore should hold for almost all wave fields. In terms of the 2-point Wigner this leads to the further condition

$$\lim_{\varepsilon\to 0} \mathbb{E}_{\mu_\varepsilon}\big(W^\varepsilon_\psi(\lfloor r_1\rfloor_\varepsilon, k_1) W^\varepsilon_\psi(\lfloor r_2\rfloor_\varepsilon, k_2)\big) = W(r_1,k_1)W(r_2,k_2) \tag{3.7}$$

provided $r_1 \neq r_2$, which in the context of rarified gas dynamics is known as the assumption of molecular chaos.

Under the assumptions (3.6) and (3.7) let us now study the time evolution generated by the flow $U(t)$. The time scale has to be adjusted to the spatial scale. Since the speed of propagation is order 1, the time scale must be $\varepsilon^{-1}\tau$ with $\tau = \mathcal{O}(1)$. A standard semiclassical analysis yields that the limit Wigner function exists and is governed by the linear transport equation

$$\frac{\partial}{\partial\tau} W(r,k,\tau) + \frac{1}{2\pi}\nabla_k\omega(k)\cdot\nabla_r W(r,k,\tau) = 0 \tag{3.8}$$

with initial conditions $W(r,k,0) = W(r,k)$. The factor $1/2\pi$ comes from our definition of the Fourier transform. Molecular chaos also holds in the sense that

$$\begin{aligned}\lim_{\varepsilon\to 0} \mathbb{E}_{\mu_\varepsilon}\big(W_{U(\tau/\varepsilon)\psi}(\lfloor r_1\rfloor_\varepsilon, k_1) W_{U(\tau/\varepsilon)\psi}(\lfloor r_2\rfloor_\varepsilon, k_2)\big)& \\ = W(r_1,k_1,\tau)W(r_2,k_2,\tau)&\end{aligned} \tag{3.9}$$

provided $r_1 - (2\pi)^{-1}\nabla\omega(k_1)\tau \neq r_2 - (2\pi)^{-1}\nabla\omega(k_2)\tau$.

If, in addition, the initial measure has good spatial mixing properties, then the Wigner function $W(r,k,\tau)$ can be given also a probabilistic interpretation on the scale of the lattice [7]. On $\mathbb{Z}^d$ one considers the reference point $\lfloor\varepsilon^{-1}r\rfloor$ and an arbitrary bounded box centered at $\lfloor\varepsilon^{-1}\rfloor$. Then the measure $\mu_\varepsilon \circ U(-\varepsilon^{-1}\tau)$ restricted to this box converges as $\varepsilon \to 0$ to a translation and gauge invariant Gaussian measure with covariance $W(r,k,\tau)$ in Fourier space, depending parametrically on r,τ.

Despite a somewhat winding discussion, I hope to have sufficiently emphasized that the free flow $U(t)$ forces the local statistics to be Gaussian, translation, and gauge invariant. The limit $\varepsilon \to 0$ is identical to the standard semiclassical limit of the Schrödinger equation. There is one important difference however which needs to be pointed out.

The semiclassical limit of the Schrödinger equation holds for deterministic initial data, i.e. for a sequence of initial wave functions ψ^ε such that the induced sequence of Wigner functions $W^\varepsilon_{\psi^\varepsilon}$ converges weakly to a limit probability measure on the classical phase space $\mathbb{R}^d \times \mathbb{T}^d$. In contrast, the kinetic limit can hold only for initial data which are sufficiently random. Analytically such assumptions are hidden in the smoothness of $W(r,k)$ and in properties like (3.7). The Gaussian measures discussed above have in a certain sense maximal randomness. How much the initial randomness can be relaxed is a little understood topic.

We return now to the case of weak nonlinearity, $\lambda > 0$, $\lambda \ll 1$. Let us first consider the spatially homogeneous case with an initial measure, μ_{G}, which is gauge invariant, Gaussian, and satisfies

$$\mathbb{E}_{\mu_{\mathrm{G}}}\big(\psi(x)\big) = 0\,, \quad \mathbb{E}_{\mu_{\mathrm{G}}}\big(\psi(x)\psi(x')\big) = 0\,, \tag{3.10}$$

for all $x, x' \in \mathbb{Z}^d$, and

$$\mathbb{E}_{\mu_{\mathrm{G}}}\big(\psi(x)^*\psi(x')\big) = \int_{\mathbb{T}^d} \mathrm{e}^{\mathrm{i}2\pi k\cdot(x-x')} W(k)\mathrm{d}k \tag{3.11}$$

with a bounded and continuous W. As explained, during the initial time slip such a measure would be established anyhow through the linear part of the dynamics. The zero averages (3.10) are preserved under the dynamics and, as a definition, the covariance function at time t is given by

$$\mathbb{E}_{\mu_{\mathrm{G}}}\big(\psi_t(x)^*\psi_t(x)\big) = \int_{\mathbb{T}^d} \mathrm{e}^{\mathrm{i}2\pi k\cdot(x-x')} W_\lambda(k,t)\mathrm{d}k\,. \tag{3.12}$$

For $\lambda = 0$, μ_{G} is invariant and $W_0(k,t) = W(k)$. The nonlinearity will drive the system away from the initial Gaussian measure, but the linear part pushes back to the "manifold" of gauge and translation invariant Gaussian measures. Thus we expect a slow evolution of $W_\lambda(k,t)$. Since the collision rate is of order λ^2, $W_\lambda(k,t)$ will vary for $t = \mathcal{O}(\lambda^{-2})$. In addition, to compute the evolution equation for W_λ we may assume that the next time step can be computed under the assumption that the measure at the current time $t = \lambda^{-2}\tau$ is on the Gaussian manifold. This reasoning leads to the

Kinetic Conjecture (spatially homogeneous). Let the initial measure μ_{G} be given as in (3.10), (3.11). Then the limit

$$\lim_{\lambda\to 0} W_\lambda(k, \lambda^{-2}\tau) = W(k,\tau) \tag{3.13}$$

exists and $W(k,\tau)$ solution of the kinetic equation

$$\frac{d}{d\tau} W(\tau) = \mathcal{C}\big(W(\tau)\big) \tag{3.14}$$

with initial data $W(0) = W$. Here $\mathcal{C}$ is the cubic Peierls-Boltzmann collision operator defined by

$$\mathcal{C}(W)(k_1) = 4\pi \int_{(\mathbb{T}^d)^3} \delta(k_1 + k_2 - k_3 - k_4)\delta(\omega_1 + \omega_2 - \omega_3 - \omega_4)|\widehat{V}(k_2 - k_3)|^2 \times (W_1W_3W_4 + W_2W_3W_4 - W_1W_2W_3 - W_1W_2W_4)\mathrm{d}k_2\mathrm{d}k_3\mathrm{d}k_4\,, \tag{3.15}$$

where $\omega_j = \omega(k_j)$, $W_j = W(k_j)$, $j = 1, 2, 3, 4$. ◇

For a detailed argument we refer to [8], where a lattice Fermi fluid is studied. Replacing in [8] the anticommuting Fermi field, $a(k)$, by the commuting field ψ and replacing "quasifree" by Gaussian, one arrives at a concise formal argument for the validity of (3.14), (3.15).

To allow for spatial variation we follow the scheme of the semiclassical limit with

$$\varepsilon = \lambda^2\,. \tag{3.16}$$

The sequence of initial measures, $\mu_{\mathrm{G},\varepsilon}$, is gauge invariant, Gaussian and satisfies (3.6) and (3.7). This means that locally the measure is approximately Gaussian and translation invariant. Therefore one can use the same reasoning as in the translation invariant case.

Kinetic Conjecture (spatial inhomogeneous). Let the sequence of initial measures $\mu_{\mathrm{G},\varepsilon}$ be gauge invariant, Gaussian and satisfying the property (3.6) and let $W_\lambda(r, k, \tau)$ be the scaled 1-point Wigner function at time $t = \lambda^{-2}\tau$ and location $\lfloor r \rfloor_\varepsilon$. Then the limit

$$\lim_{\lambda \to 0} W_\lambda(r, k, \tau) = W(r, k, \tau) \tag{3.17}$$

exists and $W(r, k, \tau)$ is the solution to the kinetic equation

$$\frac{\partial}{\partial \tau} W(r, k, \tau) + \frac{1}{2\pi} \nabla_k \omega(k) \cdot \nabla_r W(r, k, \tau) = \mathcal{C}\big(W(r, k, \tau)\big)\,. \tag{3.18}$$

Here the collision operator acts locally on the k variables, according to (3.15), at fixed r, τ. ◇

A proof of either conjecture remains open. There are two somewhat disjoint approaches. Benedetto, Castella, Esposito and Pulvirenti [3], see also the previous ICM contribution of Pulvirenti [15], study the weak coupling limit for quantum gases. Their methods and results can be translated to the present case, although this has never been written out in complete detail. They start from the hierarchy of multi-point Wigner functions. The free part leaves the space of n-point Wigner functions invariant, while the nonlinearity couples n to $n + 1$. Thus it is natural to expand in λ. While the series cannot be controlled, one can study each term separately. This program is followed up in [3] in their

context, which corresponds to the nonlinear Schrödinger equation (1.1) on $\mathbb{R}^d$ and with initial data of slow spatial variation. As an alternative approach [13] we developed the Duhamel expansion for (2.7), which is the basic strategy to prove the kinetic limit of equilibrium time correlations, see Section 4 below.

As a byproduct our proof provides also some information on the kinetic conjecture in the spatially homogeneous case. We assume $d \geq 4$, $\widehat{V}(k) = 1$, and implicit conditions on ω, which are established to hold for the lattice Laplacian. Let us first define the collision operator $\mathcal{C}_{j,n+2}$. $C_{j,n+2}$ acts on functions of $k_1, \ldots, k_{n+2}$, but only hitting the arguments k_j, k_{n+1}, k_{n+2}. For these three variables we use (3.15) with the integration over k_4 worked out explicitly by using the momentum conservation $\delta(k_1 + k_2 - k_3 - k_4)$. We also define

$$\mathcal{C}_{n+2} = \sum_{j=1}^{n} \mathcal{C}_{j,n+2} \,. \tag{3.19}$$

Theorem 3.1. *Let the initial measure, μ_{G}, satisfy (3.10) and (3.11) with continuous W. Then the odd terms of the Duhamel expansion vanish as $\lambda \to 0$. For the term of order $2n$, $n \in \mathbb{N}$, of the Duhamel expansion the limit $\lambda \to 0$ exists and is given by*

$$\frac{\tau^n}{n!}(\mathcal{C}_3 \ldots \mathcal{C}_{2n+1}\widehat{\rho}_{2n+1})(k_1) \tag{3.20}$$

with

$$\widehat{\rho}_{2n+1}(k_1, \ldots, k_{2n+1}) = \prod_{j=1}^{2n+1} W(k_j) \,. \tag{3.21}$$

Note that (3.20) is the n-th order Taylor expansion for the solution of kinetic equation (3.14) at $\tau = 0$.

Taking absolute values and counting the number of terms, the expression (3.20) is bounded by

$$\tau^n \frac{1}{n!} c^{2n+1}(1 \cdot 3 \cdot \ldots \cdot (2n+1)) \,, \tag{3.22}$$

hence having a finite radius of convergence for the sum over n. The naive bound on the Duhamel expansion has a further factor of $n!$ coming from the initial Gaussian measure. To fight the zero radius of convergence one cuts the series at some large, λ-dependent n. But then one needs a priori estimate on the solution to handle the remainder. Such a property is badly missing. In the case of the linear Schrödinger equation with a weak random potential one can use that $\|\psi(t)\|$ is conserved [9]. While still true for the nonlinear Schrödinger equation, this by itself does not suffice for a workable a priori bound.

4. Equilibrium Time Correlations

Our enterprise started from the elementary observation that for equilibrium time correlations one can use Schwarz inequality and time stationarity to bound

the error term of the Duhamel expansion. To actually control the series still requires a substantial effort [13]. But for the first time one controls the kinetic limit for a weakly nonlinear wave equation, at least for initial conditions which are local perturbations away from equilibrium.

We start from the finite volume Gibbs measure

$$\frac{1}{Z_\Lambda} \exp\big[-\beta\big(H(\psi)-\lambda_0 N(\psi)\big)\big] \prod_{x\in\Lambda} \big(\mathrm{d}\Re\psi(x)\mathrm{d}\Im\psi(x)\big) = \mu_\Lambda^\lambda\,. \tag{4.1}$$

Here $\Lambda \subset \mathbb{Z}^d$ is some box, $\beta > 0$ the inverse temperature, λ_0 the chemical potential, and the partition function Z_Λ makes μ_Λ^λ a probability measure. Expectations with respect to μ_Λ^λ are denote by $\mathbb{E}_\Lambda^\lambda$. For $\lambda = 0$, (4.1) is a Gaussian measure. Its infinite volume limit has the covariance

$$\mathbb{E}^0\big(\psi(x)^*\psi(x')\big) = \int_{\mathbb{T}^d} \mathrm{e}^{\mathrm{i}2\pi k\cdot(x-x')} W^{\mathrm{eq}}(k)\mathrm{d}k \tag{4.2}$$

with the equilibrium covariance function

$$W^{\mathrm{eq}}(k) = \big(\beta(\omega(k)-\lambda_0)\big)^{-1}\,. \tag{4.3}$$

Thus one has to impose the condition

$$\omega(k) > \lambda_0 \quad \text{all } k \in \mathbb{T}^d\,. \tag{4.4}$$

Then the Gaussian measure defined by (4.2) has exponential decay of correlations. If $\omega(0) = \lambda_0$, and otherwise (4.4) holds, the measure in (4.2) has slow decay of correlations and on top there could be a condensate component [4], see also references in [20]. While this case is of great current interest physically, a mathematical proof of the kinetic limit is out of reach, presently.

In a recent contribution [1], it is proved that the above properties remain intact for sufficiently small λ. In particular, the infinite volume limit, $\Lambda \uparrow \mathbb{Z}^d$, of the sequence of measures in (4.1) exists and defines a unique Gibbs measure, μ^λ, on $\mathbb{C}^{\mathbb{Z}^d}$. Expectations with respect to μ^λ are denoted by $\mathbb{E}^\lambda$. Of importance in our context is a sharp estimate of the closeness to the $\lambda = 0$ Gaussian measure.

Theorem 4.1. *(i) Let $\beta > 0$ and (4.4) hold. Then, for all $0 < \lambda < \overline{\lambda}$ with a sufficiently small $\overline{\lambda}$, the fully truncated correlation functions (cumulants) have the bound*

$$\sup_{\Lambda,\sigma\in\{\pm1\}^n} \sum_{x\in\Lambda^n} \delta_{x_1,0} \left| \mathbb{E}_\Lambda^\lambda \left(\prod_{j=1}^n \psi(x_j,\sigma_j) \right)_{\mathrm{T}} \right| \le \lambda (c_0)^n n! \tag{4.5}$$

for all $n \ge 4$, where $\sigma = (\sigma_1,\dots,\sigma_n)$, $x = (x_1,\dots,x_n)$.

(ii) For the two-point function it holds, with $\Lambda = [-L/2, L/2]^d$,

$$\limsup_{L\to\infty} \sum_{|x|<L/2} \big|\mathbb{E}_\Lambda^\lambda\big(\psi(0)^*\psi(x)\big) - \mathbb{E}^0\big(\psi(0)^*\psi(x)\big)\big| \le 2\lambda(c_0)^2\,. \tag{4.6}$$

The most basic time correlation is the two-point function

$$\mathbb{E}^{\lambda}\big(\psi_t(x)^*\psi_0(x')\big) = C_\lambda(x-x',t)\,. \tag{4.7}$$

At $\lambda = 0$ it is purely oscillatory in Fourier space,

$$\widehat{C}_0(k,t) = \big(\beta(\omega(k)-\lambda_0)\big)^{-1}\mathrm{e}^{\mathrm{i}\omega(k)t}\,. \tag{4.8}$$

For $\lambda > 0$, but $\lambda \ll 1$, one expects these oscillations to be damped. We will prove that this indeed happens on the time scale $\lambda^{-2}\tau$.

Our proof is structured in such a way that the dispersion relation ω is kept general. There will be implicit conditions imposed on ω. The simplest one is the ℓ_3-*dispersivity* which states that, with the definition

$$p_t(x) = \int_{\mathbb{T}^d} \mathrm{e}^{-\mathrm{i}\omega(k)t}\mathrm{e}^{\mathrm{i}2\pi k\cdot x}\mathrm{d}k\,, \quad x\in\mathbb{Z}^d\,, \tag{4.9}$$

one has for some $\delta > 0$

$$(\|p_t\|_3)^3 \le c(1+|t|)^{-(1+\delta)}\,. \tag{4.10}$$

(4.10) can be proved by stationary phase methods and holds generically for dimension $d \ge 3$. In addition we require the *constructive interference bound*, which involves a d-dimensional oscillatory integral, and the *crossing bounds*, which involve $2d$-dimensional oscillatory integrals. These conditions are somewhat technical to state and we refer to [13] for their precise formulation.

We prove our conditions for the lattice Laplacian. The interference bound requires then $d \ge 4$. It would be of interest to study these bounds for a more general class of dispersion relations. For sure, the condition (4.10) will stay. Thus $d \ge 3$ seems to be necessary for the existence of the kinetic limit. For the constructive interference bound there is some freedom and it is conceivable that an improved version of the proof will also cover $d = 3$.

From a first order expansion in λ, one infers that already at time scale $\lambda^{-1}\tau$, which is short compared to the dissipative scale $\lambda^{-2}\tau$, there are additional oscillatory contributions. More systematically they can be understood by renormalizing $\omega(k)$ to

$$\omega^{\lambda}(k) = \omega(k) + \lambda\int_{\mathbb{T}^d}\big(\widehat{V}(0)+\widehat{V}(k-k_1)\big)W^{\mathrm{eq}}(k_1)\mathrm{d}k_1\,. \tag{4.11}$$

There will be further frequency shifts of order λ^2. The renormalization (4.11) plays a distinguished role, since it must be included in the free part and should not be Duhamel expanded. Clearly for $\widehat{V}(k) = 1$, the order λ correction in (4.11) is independent of k, which is a welcome simplification. We believe that our proof works also beyond the case $V(x) = \delta_{x,0}$ without substantial changes. This is a further point which we will have to address in the future.

To state our result we still have to define the dissipative terms. With real Γ_1 and Γ_2 let

$$\Gamma(k_1) = \Gamma_1(k_1) + \mathrm{i}\Gamma_2(k_1) = -2\int_0^\infty \mathrm{d}t \int_{(\mathbb{T}^d)^3} \mathrm{d}k_2\mathrm{d}k_3\mathrm{d}k_4 \delta(k_1+k_2-k_3-k_4) \times \mathrm{e}^{\mathrm{i}t(\omega_1+\omega_2-\omega_3-\omega_4)}(W_3^{\mathrm{eq}}W_4^{\mathrm{eq}} - W_2^{\mathrm{eq}}W_4^{\mathrm{eq}} - W_2^{\mathrm{eq}}W_3^{\mathrm{eq}}) \tag{4.12}$$

and note that, by explicit computation,

$$\Gamma_1(k_1) = 2\pi W^{\mathrm{eq}}(k_1)^{-2} \int_{(\mathbb{T}^3)^d} \mathrm{d}k_2\mathrm{d}k_3\mathrm{d}k_4 \delta(k_1+k_2-k_3-k_4) \times \delta(\omega_1+\omega_2-\omega_3-\omega_4) \prod_{j=1}^4 W^{\mathrm{eq}}(k_j) \geq 0\,. \tag{4.13}$$

The ℓ_3-bound (4.10) ensures the existence of the integrations in (4.12) and (4.13).

Theorem 4.2. *Let $d \geq 4$, $\widehat{V} = 1$, and α the nearest neighbor Laplacian. Then there exists $t_0 > 0$ such that for all $|t| < t_0$ it holds*

$$\lim_{\lambda\to 0} \mathbb{E}^\lambda\big(\langle \widehat{f}, \widehat{\psi}_0\rangle^* \langle \mathrm{e}^{-\mathrm{i}\omega^\lambda \lambda^{-2}t}\widehat{g}, \widehat{\psi}_{\lambda^{-2}t}\rangle\big) = \int_{\mathbb{T}^d} \widehat{f}(k)^*\widehat{g}(k)\mathrm{e}^{-\Gamma_1(k)|t| - \mathrm{i}t\Gamma_2(k)} W^{\mathrm{eq}}(k)\mathrm{d}k\,. \tag{4.14}$$

The restriction to the finite kinetic t_0 is hidden in the finite radius of convergence for the sum over n in (3.20). Very roughly, in the Schwarz bound for the error term in the Duhamel expansion, there are many terms which vanish as $\lambda \to 0$. However there are still the terms of (3.20) which are not zero. Schematically the remainder is then bounded by

$$\sum_{n=n_0(\lambda)}^\infty |t/t_0|^n \tag{4.15}$$

with $n_0(\lambda) \to \infty$ as $\lambda \to 0$. For (4.15) to vanish as $\lambda \to 0$ requires $|t/t_0| < 1$.

5. Fluctuation Field

Under the equilibrium measure, $\psi_t(x)$ is a stochastic process stationary in $t \in \mathbb{R}$ and $x \in \mathbb{Z}^d$. For small λ, $\psi_t(x)$ has rapid oscillations which can be subtracted by defining

$$\langle f, \phi_t^\varepsilon\rangle = \int_{\mathbb{T}^d} \widehat{f}(k)^* \mathrm{e}^{\mathrm{i}\omega^\lambda(k)t/\varepsilon}\widehat{\psi}_{t/\varepsilon}(k)\mathrm{d}k\,, \quad \varepsilon = \lambda^2\,, \tag{5.1}$$

with $f \in \ell_2(\mathbb{Z}^d)$. Note that only time is rescaled while space euqals $\mathbb{Z}^d$ independent of ε. Our main Theorem 4.2 can be rephrased that the covariance of $\phi_t^\varepsilon(x)$ has a limit as $\varepsilon \to 0$. Thus it is natural to ask whether the full process has a limit.

We discuss first the limit process, denoted by $\phi_t(x)$. It is a gauge invariant Gaussian process with covariance

$$\mathbb{E}\big(\langle f, \phi_t\rangle\langle \phi_{t'}, g\rangle\big) = \int_{\mathbb{T}^d} \widehat{f}(k)^* \widehat{g}(k) \mathrm{e}^{-\Gamma_1(k)|t-t'| - \mathrm{i}(t-t')\Gamma_2(k)} W^{\mathrm{eq}}(k) \mathrm{d}k\,. \tag{5.2}$$

(5.2) defines an infinite dimensional Ornstein-Uhlenbeck process, governed by

$$\mathrm{d}\phi_t(x) = A\phi_t(x)\mathrm{d}t + \mathrm{d}\eta(x,t)\,, \quad x \in \mathbb{Z}^d\,. \tag{5.3}$$

The linear operator A is convolution with the Fourier inverse of $-\Gamma$, $\Gamma = \Gamma_1 + \mathrm{i}\Gamma_2$. $\eta(x,t)$ is complex valued Brownian motion in t and has a spatial covariance given by the Fourier inverse of $2\Gamma_1(k)W^{\mathrm{eq}}(k)$. The invariant measure of the Ornstein-Uhlenbeck process has covariance W^{eq} in Fourier space and the process is time reversible meaning that

$$\phi_t = \phi_{-t}^* \tag{5.4}$$

in probability.

The Duhamel expansion can deal only with moments of the type

$$\mathbb{E}^\lambda \left(\prod_{j=1}^n \langle f_j, \phi_{t_j}^\varepsilon\rangle \langle \phi_{t_{j+n}}^\varepsilon, g_j\rangle \right) \tag{5.5}$$

for $f_j, g_j \in \ell_2(\mathbb{Z}^d)$ and for arbitrary times $t_1, \ldots, t_{2n}$. There seems to be no simple trick through which the moments (5.5) could be reduced to the covariance. Rather one has to work out the Duhamel expansion for the product appearing (5.5) and use Hölder inequality and stationary to bound the error term in the expansion. While the details have not been written out, at the level of the 4-th moment our method works fine and presumably also for higher moments. Thus under the conditions of Theorem 4.2, in particular for $|t_j| < t_0$, $j = 1, \ldots, 2n$, it holds that

$$\lim_{\varepsilon \to 0} \langle f, \phi_t^\varepsilon\rangle = \langle f, \phi_t\rangle\,, \tag{5.6}$$

where the convergence is in the sense of convergence of the moments (5.5).

There are correlations of physical interest which are not covered by (5.6). An example may suffice at this stage. The local energy current at site x at time t is defined by

$$\mathcal{J}_\alpha(x,t) = \sum_{y \in \mathbb{Z}^d} \big(\psi(x)^* \gamma_\alpha(x-y)\psi(y) + \psi(y)^* \gamma_\alpha(x-y)\psi(x)\big) \tag{5.7}$$

with $\widehat{\gamma}_\alpha(k) = \big(\nabla_\alpha \omega(k)\big)\omega(k)$, $\alpha = 1, \dots, d$. Energy dissipation can be characterized by the total energy current correlation, which is defined through

$$\sum_{x\in\mathbb{Z}^d} \mathbb{E}^\lambda\big(\mathcal{J}_\alpha(x,t)\mathcal{J}_{\alpha'}(0,0)\big) = C^{\mathrm{cur}}_{\lambda,\alpha\alpha'}(t)\,, \tag{5.8}$$

using that $\mathbb{E}^\lambda\big(\mathcal{J}_\alpha(x,t)\big) = 0$. As before one would like to establish the limit $\lambda \to 0$ of

$$C^{\mathrm{cur}}_{\lambda,\alpha\alpha'}(\lambda^{-2}t)\,. \tag{5.9}$$

Because of the sum over x in (5.8) this limit is not covered by (5.6). In fact, an interchange of limit and sum is not expected to be valid. To establish the limit in (5.9) is a future challenge of interest.

References

[1] A. Abdesselam, A. Procacci, B. Scoppola, *Clustering bound on n-point correlations for unbounded spin systems*, arXiv:0901.4756, preprint (2009).

[2] G. Bal, *Lecture Notes: Waves in Random Media*, www.columbia.edu/∼g62030/COURSES/E6901-Waves/Lecture-Waves.pdf.

[3] D. Benedetto, F. Castella, R. Esposito, M. Pulvirenti, *From the N-body Schrödinger equation to the quantum Boltzmann equation, a term-by-term convergence result in the weak coupling regime*, Comm. Math. Phys. **277** (2008), 1–44.

[4] T. Berlin, M. Kac, *The spherical model of a ferromagnet*, Phys. Rev. **86** (1952), 821–835.

[5] P. Buttà, E. Caglioti, S. Di Ruzza, C. Marchioro, *On the propagation of a perturbation in an anharmonic crystal*, J. Stat. Phys. **127** (2007), 313–325.

[6] T.V. Dudnikova, A.I. Komech, H. Spohn, *On the convergence to statistical equilibrium for harmonic crystals*, J. Math. Phys. **44** (2003), 2596–2620.

[7] T.V. Dudnikova, H. Spohn, *Local stationarity for lattice dynamics in the harmonic approximation*, Markov Processes Rel. Fields **12** (2006), 645–678.

[8] L. Erdös, M. Salmhofer, H.T. Yau, *On the quantum Boltzmann equation*, J. Stat. Phys. **116** (2004), 367–380.

[9] L. Erdös, H.T. Yau, *Linear Boltzmann equation as the weak coupling limit of a random Schrödinger equation*, Comm. Pure Appl. Math. **53** (2000), 667–735.

[10] J.L. van Hemmen, *Dynamics and ergodicity of the infinite harmonic crystal*, Ph.D. Thesis, Univ. Groningen, 1976.

[11] T.G. Ho, L.J. Landau, *Fermi gas on a lattice in the van Hove limit*, J. Stat. Phys. **87** (1997), 821–845.

[12] O.E. Lanford, J.L. Lebowitz, E.H. Lieb, *Time evolution of infinite harmonic systems*, J. Stat. Phys. **16** (1977), 453–461.

[13] J. Lukkarinen, H. Spohn, *Weakly nonlinear Schrödinger equation with random initial data*, arXiv:math-ph/0901.3283, preprint (2009).

[14] R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford University Press, Oxford 1989.

[15] M. Pulvirenti, *The weak-coupling limit for large classical and quantum systems*, Proceedings, ICM 2006 at Madrid, Vol. III, p. 229–256.

[16] D. Ruelle, *Chance and Chaos*, Princeton Science Library, Princeton University Press, Princeton, 1993.

[17] L. Ryzhik, G. Papanicolaou, J.B. Keller, *Transport equations for elastic and other waves in random media*, Wave Motion **24** (1996), 327–370.

[18] B. Simon, *The $P(\phi)_2$ Euclidean (Quantum) Field Theory*, Princeton University Press, Princeton, 1974.

[19] H. Spohn, *The phonon Boltzmann equation, properties and link to weakly anharmonic lattice dynamics*, J. Stat. Phys. **124** (2006), 1041–1104.

[20] H. Spohn, *Kinetics of the Bose-Einstein condensation*, Physica D, online, PHYS D 30806.

[21] V.E. Zakharov, V.S. L'vov, G. Falkovich, *Kolmogorov Spectra of Turbulence I: Wave Turbulence*, Springer, Berlin, 1992.

Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

On the Geometry of Singularities in Quantum Field Theory

Katrin Wendland*

Abstract

This survey investigates the geometry of singularities from the viewpoint of conformal and topological quantum field theory and string theory.

First, some classical results concerning simple surface singularities are collected, paying special attention to the ubiquitous ADE theme. For conformal field theory, recent progress both on axiomatic and on constructive issues is discussed, as well as a well established classification result, which is also related to the ADE theme, but not complete. Special focus concerning constructive results is owed to superconformal field theories associated to $K3$ surfaces and some of their higher dimensional cousins. Finally, for topological quantum field theories, their role between conformal field theory and singularity theory is reviewed, along with the origin of tt^* geometry, and some of its applications.

Mathematics Subject Classification (2010). Primary 14E15; Secondary 14E16, 14J28, 17B68, 32S30, 32S45, 81T40, 81T45.

Keywords. Conformal field theory; topological field theory; singularity theory.

Introduction

Ἡ φύσις οὐδέν ποιεῖ ἅλματα - *Nature does not make jumps* - or does she?

Catastrophes and other phenomena related to singularity theory have challenged this Aristotelean thesis at least since the days of Felix Klein in the 1880s, and quantum phenomena in general are hardly compatible with it. On the other hand, the description of singularities often naturally amounts to a detailed study of a smooth neighborhood of a singularity, and in quantum physics, classical and smooth limits are of paramount importance. Perhaps it is fair to

*Lehrstuhl für Analysis und Geometrie
Institut für Mathematik, Universität Augsburg, Universitätsstr. 14, D-86159 Augsburg, Germany. E-mail: katrin.wendland@math.uni-augsburg.de.

say that nowadays, the dialectic between smooth and singular phenomena is a driving force in geometry and in quantum physics. This, roughly, is the theme of the present note.

It is hardly possible to give a complete overview on this broad topic, let alone to give credit to all the contributors who deserve it. Therefore, this survey gives my very personal account, attempting to address those aspects of singularity theory which from my subjective point of view have influenced recent progress in mathematical physics, as well as those which are certain to continue to be of relevance. My personal interest lies in the interactions between geometry and quantum field theory. And of course, putting false modesty aside, I take this opportunity to summarize some of my collaborators' and my own more recent results and ideas in the area. This includes some progress in our understanding of conformal field theory, both from an axiomatic and a constructive viewpoint. On the other hand, topological quantum field theories and specifically their geometric content and classification are addressed. The overall theme remains the role of singularities and their geometry in various areas of quantum field theory.

This note is structured as follows:

The following Section 1 gives a brief overview on some basic well established notions from singularity theory, with a particular focus on simple singularities, including their deformation theory in terms of unfoldings, as well as their minimal resolutions. As a welcome necessity, the theme of ADE classifications is predominant. Moreover, I review Kodaira's classification of degenerate fibers in elliptic fibrations of complex surfaces. Finally, various aspects of the McKay correspondence are summarized, including the special McKay correspondence which quite undeservingly is little known, at least to physicists.

Section 2 is devoted to conformal and superconformal field theory. The exposition includes a rough definition of conformal field theory, which has been unpublished in this form, so far, with details to be presented elsewhere [83]. In addition, better established notions are discussed, including that of the $N = 2$ super-Virasoro algebra and the conformal field theoretic elliptic genus. I then turn to superconformal field theories associated to $K3$ and their geometric interpretations. In particular, having the main theme "singularity theory" in mind, orbifold conformal field theories associated to $K3$ are discussed. Here, the determination of the B-field values on the exceptional divisor of the resolution of quotient singularities is of interest, as well as their interpretation in terms of the classical McKay correspondence. Moreover, the geometric interpretation of the twist fields in orbifold conformal field theories is briefly addressed. Finally, non-classical dualities are presented, including a version of mirror symmetry for elliptically fibered $K3$ surfaces. Another non-classical duality allows for the construction of superconformal field theories associated to certain smooth quartic $K3$ surfaces.

Section 3 summarizes some basic concepts in topological field theory. In particular, the role of singularity theory for these quantum field theories is

addressed, explaining how the notions of Frobenius algebras, Frobenius manifolds, and tt^* geometry arise naturally in this context. Since already in the 1990s the proof of integrability of the tt^* equations led to yet another ADE classification, namely that of the $N = (2, 2)$ Virasoro minimal models with spacetime supersymmetry, I briefly review these results and comment on the classification of all $N = (2, 2)$ Virasoro minimal models, which has not been completed, so far.

This note ends with an outlook on work in progress in the context of singularities in threefolds in Section 4; the construction of superconformal field theories associated to so-called Borcea-Voisin threefolds as well as ongoing investigations of elliptically fibered Calabi-Yau threefolds are addressed.

1. Some Background on Singularities

This section is devoted to a summary of some well known but fundamental mathematical background, concerning singularities, their resolution, and their deformation. The literature on this topic is vast; standard references are the classical textbooks by Arnol'd, Gusein-Zade and Varchenko [3]; a very nice introductory overview by Brieskorn can be found in [15]; a less basic and more extensive review is Kollár's Seattle lecture on the topic [55].

In very general terms, a SINGULARITY is a point $x \in X$ in a variety X such that X is not smooth in x. Both the classification and desingularization of such singularities have been a driving force in algebraic geometry, at least since the days of Riemann. In modern terminology, a RESOLUTION of X is a smooth variety $\widetilde{X}$ together with a projective birational morphism $\pi\colon \widetilde{X} \to X$. Probably the most influential paper on the topic is Heisuke Hironaka's seminal work [47], containing an existence proof for resolutions of singularities of varieties in arbitrary dimensions over fields of characteristic zero by repeated blow-up along non-singular subvarieties. In the following, I always work in characteristic zero and more specifically with varieties over the field of complex numbers $\mathbb{C}$. Moreover, unless stated otherwise, I discuss special classes of ISOLATED SINGULARITIES.

The neighborhood of a singular point in an n-dimensional variety can often be described as the zero set of a holomorphic function germ $f\colon (\mathbb{C}^{n+1}, 0) \to (\mathbb{C}, 0)$ with an isolated critical point at 0. We then speak of a HYPERSURFACE SINGULARITY. Let me introduce some of the standard notions describing singularities of this type:

Definition 1.1. Let f denote a holomorphic function germ $f\colon (\mathbb{C}^{n+1}, 0) \to (\mathbb{C}, 0)$ with an isolated singularity at 0.

1. The JACOBI ALGEBRA OF f is the algebra

$$\mathcal{J} := \mathbb{C}[x_0, \ldots, x_n] / \left(\frac{\partial f}{\partial x_0}, \ldots, \frac{\partial f}{\partial x_n} \right).$$

The dimension μ of $\mathcal{J}$ is the MILNOR NUMBER of the singularity.

2. A function germ $\widetilde{f}\colon (\mathbb{C}^{m+1},0)\to(\mathbb{C},0)$ with an isolated singularity at 0 is STABLY EQUIVALENT TO f if after addition of quadratic forms in an appropriate number of additional variables, f and $\widetilde{f}$ are related by a diffeomorphic change of independent variables.

3. A SEMIUNIVERSAL UNFOLDING of the singularity given by f is a holomorphic function germ $F\colon (\mathbb{C}^{n+1}\times\mathbb{C}^{\mu},0)\to(\mathbb{C},0)$, such that

$$\text{for } (x;t)=(x_0,\dots,x_n;t_1,\dots,t_\mu)\in\mathbb{C}^{n+1}\times\mathbb{C}^{\mu},\quad f(x)=F(x;t=0),$$

and the partial derivatives $\frac{\partial F}{\partial t_i}(x;t=0)$, $i\in\{1,\dots,\mu\}$, represent a basis of the Jacobi algebra $\mathcal{J}$ of f.

4. The singularity given by f is SIMPLE if for every semiuniversal unfolding $F\colon (\mathbb{C}^{n+1}\times\mathbb{C}^{\mu},0)\to(\mathbb{C},0)$ of f with $F_t(x):=F(x;t)$ for $(x;t)\in\mathbb{C}^{n+1}\times\mathbb{C}^{\mu}$, in a sufficiently small neighbourhood $M\subset\mathbb{C}^{\mu}$ of $0\in\mathbb{C}^{\mu}$ the fibers $F_t^{-1}(0)$, $t\in M$, exhibit only a finite number of pairwise stably non-equivalent singularities. The space M is called the BASE of the semiuniversal unfolding.

According to a fundamental result by Arnol'd [1, 2], the germs of holomorphic functions with simple singularities enjoy an ADE-CLASSIFICATION:

Theorem 1.2. *If $f\colon(\mathbb{C}^{n+1},0)\to(\mathbb{C},0)$ is the holomorphic function germ of a simple singularity, then f is stably equivalent to one of the following:*

$$\begin{aligned}
&A_k\colon f(x,y)=x^2+y^{k+1}, && k\geq 1,\\
&D_k\colon f(x,y)=x^2y+y^{k-1}, && k\geq 4,\\
&E_6\colon f(x,y)=x^3+y^4,\\
&E_7\colon f(x,y)=x^3+xy^3,\\
&E_8\colon f(x,y)=x^3+y^5.
\end{aligned}$$

In other words, the singularity is stably equivalent to one of the quotient surface singularities studied by Schwarz [72] *and Klein* [51, 52]*, namely $0\in\mathbb{C}^2/\Gamma$ with $\Gamma\subset SU(2)$ a finite subgroup:*

$$\begin{aligned}
A_k\colon &\quad \text{cyclic group of order } k+1,\\
D_k\colon &\quad \text{binary dihedral group of order } 4(k-2),\\
E_6\colon &\quad \text{binary tetrahedral group of order } 24,\\
E_7\colon &\quad \text{binary octahedral group of order } 48,\\
E_8\colon &\quad \text{binary icosahedral group of order } 120.
\end{aligned}$$

The respective groups Γ are called the ADE TYPE FINITE GROUPS *in the following.*

By a seminal result due to Milnor [60], the topology of every isolated hypersurface singularity $f\colon (\mathbb{C}^{n+1},0) \to (\mathbb{C},0)$ with polynomial f is uniquely determined by its MILNOR FIBRATION. Here, one uses a disc $\Delta \subset \mathbb{C}$ with $0 \in \Delta$ which is small compared to the radius ϵ of a chosen ball $B_\epsilon \subset \mathbb{C}^{n+1}$ around the singular point 0 of $f^{-1}(0)$. Let $X_0 := B_\epsilon \cap f^{-1}(0)$ and $X := B_\epsilon \cap f^{-1}(\Delta)$. Then $f\colon X - X_0 \to \Delta - \{0\}$ is a locally trivial fiber bundle, the MILNOR FIBRATION. The results of [60, 13, 31, 56] yield:

Theorem 1.3. *Every fiber of the Milnor fibration of a polynomial isolated hypersurface singularity $f\colon (\mathbb{C}^{n+1},0) \to (\mathbb{C},0)$ is homotopy equivalent to a bouquet of μ real n-spheres $\mathbb{S}^n$, where μ is the Milnor number of the singularity.*

One can choose these spheres as VANISHING CYCLES *$S_1, \ldots, S_\mu$ in the sense of Lefschetz, that is, each S_i can be homotopically deformed over a path in Δ into the singular central fiber, where it vanishes. Then the intersection matrix $((S_i \cdot S_j)_{ij})$ completely determines the topology of the Milnor fibration.*

For a semiuniversal unfolding F of an isolated hypersurface singularity f as above, the vanishing cycles of Thm. 1.3 yield the LEFSCHETZ THIMBLES [67, 68]: With notations as in Def. 1.1 and the discussion before Thm. 1.3, consider $\mathcal{X} := F^{-1}(\Delta) \cap (B_\epsilon \times M) \subset \mathbb{C}^{n+1} \times \mathbb{C}^\mu$. Then the map

$$\varphi\colon \mathcal{X} \to \Delta \times M, \qquad (x;t) \mapsto (F(x;t),t) \tag{1.1}$$

is a C^∞-fibration of Milnor fibers outside a discriminant in $\Delta \times M$. A Lefschetz thimble in $\mathcal{X}_t$ then, roughly, is given as the union of a continuous family of vanishing cycles over a path in Δ which connects a critical value of F_t with a non-critical one in the boundary of Δ.

The results of Thm. 1.3 can be evaluated for the simple singularities classified in Thm. 1.2 by calculating the intersection matrices for these singularities [36, 43, 44]:

Theorem 1.4. *Consider a simple singularity $f\colon (\mathbb{C}^3,0) \to (\mathbb{C},0)$ as in Thm.* 1.2*. Let $S_1, \ldots, S_\mu$ denote vanishing cycles which generate the homology of the Milnor fibration. Then these cycles can be ordered such that the intersection matrix $((S_i \cdot S_j)_{ij})$ is the negative of the Cartan matrix of a Lie algebra of type A_k, D_k, E_6, E_7, or E_8 if and only if the singularity, according to the classification of Thm.* 1.2*, is of type A_k, D_k, E_6, E_7, or E_8, respectively.*

In view of Thm. 1.4 it is natural to expect that the universal unfolding F of a simple singularity f is governed by the data of the associated simply laced Lie algebra. The detailed relation, however, is quite subtle and has been uncovered thanks to many separate contributions, probably beginning with two famous conjectures by Grothendieck, see [75], which were proved by Brieskorn, see [14]; the excellent book [74] is devoted to a detailed exposition. Roughly, the base M of every semiuniversal unfolding of f can be identified with an open subset of the Cartan algebra of the Lie algebra associated to the singularity. In particular, there is a one-to-one correspondence between positive roots of the

Lie algebra and vanishing cycles in the Milnor fibration of the singularity. A choice of primitive roots corresponds to a choice of generators of the homology in terms of vanishing cycles as in Thm. 1.4. A vanishing cycle corresponding to a root α remains contracted to a point in the deformation F_t of f with $t \in M$ if and only if t, viewed as an element of the Cartan algebra, vanishes on α.

Apart from the beginning remarks, I have focussed the above discussion entirely on the deformation of singularities. However, according to Hironaka's result [47], all the classes of singularities discussed so far allow a resolution by repeated blow-up. Thm. 1.2 guarantees that all simple singularities can be represented as quotient singularities of the form $\mathbb{C}^2/\Gamma$ for appropriate groups Γ. For these singularities, a unique minimal resolution always exists, and the intersection matrix of the irreducible components of its exceptional divisor recovers data from the associated simply laced Lie algebra:

Theorem 1.5. *Consider a simple singularity $0 \in \mathbb{C}^2/\Gamma$ with $\Gamma \subset SU(2)$ as in Thm.* 1.2 *of type A_k, D_k, E_6, E_7, or E_8. Let $\pi\colon \widetilde{X} \to \mathbb{C}^2/\Gamma$ denote the minimal resolution of this singularity. Then the exceptional divisor of $\widetilde{X}$ is a collection of rational curves, such that every non-trivial pairwise intersection is transversal, and such that the resulting intersection matrix is the negative of the Cartan matrix of the corresponding simply laced Lie algebra of type A_k, D_k, E_6, E_7, or E_8.*

The theorem follows from work of du Val [32] and Artin [4]; the resolutions were explicitly calculated by Brieskorn [12]. A noteworthy conclusion following from a comparison between Thm. 1.4 and Thm. 1.5 is a correspondence, by means of the associated simply laced Lie algebras, between the resolution and the deformation of simple singularities.

The ADE classification of simple singularities reoccurs in a similar form for complex surfaces in the context of ELLIPTIC FIBRATIONS $p\colon X \to \Delta$. Here, without loss of generality $\Delta \subset \mathbb{C}$ is a disc containing $0 \in \mathbb{C}$, X is a smooth complex surface, and p is a proper, connected, holomorphic map, such that all fibers X_s for $s \in \Delta$ with $s \neq 0$ are elliptic. Moreover, p is assumed to be (relatively) minimal, that is, all fibers are free of (-1)-curves. Then, the following KODAIRA CLASSIFICATION holds [54]:

Theorem 1.6. *The fiber X_0 of an elliptic fibration $p\colon X \to \Delta$ as above can be of one of the following types:*

1. *If X_0 is irreducible, then it is either smooth elliptic (of so-called Kodaira type I_0), or rational with a node (of so-called Kodaira type I_1), or rational with a cusp (of so-called Kodaira type II).*

2. *If X_0 is reducible but not multiple, then X_0 is a collection of rational curves whose intersection matrix is either given by the Cartan matrix of the extended Dynkin diagram of a simply laced Lie algebra, or by a degenerate form of it. In the case of extended Dynkin diagrams, $\widetilde{A}_k$ corresponds*

to fibers of Kodaira type I_{k+1}, $\widetilde{D}_k$ corresponds to Kodaira type I^*_{k-4}, and $\widetilde{E}_6$, $\widetilde{E}_7$, $\widetilde{E}_8$ correspond to Kodaira types IV^*, III^*, II^*, respectively. Otherwise, two rational curves intersecting transversally in a double point correspond to a degenerate version of $\widetilde{A}_1$, giving Kodaira type III, while three rational curves intersecting in one point correspond to a degenerate version of $\widetilde{A}_2$, giving Kodaira type IV.

3. *If X_0 is a multiple fiber, then it is a multiple of a fiber of Kodaira type I_k.*

The result can be illustrated by observing [61] that an elliptic fibration $p\colon X \to \Delta$ with section can always be obtained by resolving the singularities of a singular Weierstraß form

$$y^2z = x^3 + a(s)xz^2 + b(s)z^3, \quad (x,y,z) \in \mathbb{CP}^2,\ s \in \Delta,\ a,b \in C^\infty(\Delta).$$

If the fiber over $s = 0$ is not smooth elliptic, then it is a rational curve with a cusp (i.e. of Kodaira type II) or with a simple singularity. Resolution, in the latter case, by Thm. 1.5 yields a fiber containing an ADE-type configuration from the exceptional divisor of the resolution. Furthermore, there is the strict transform of the original rational curve, which is responsible for the fact that Kodaira's classification yields the extended Dynkin diagrams (or a degenerate form of them) rather than the ordinary ones. This is the essence of Tate's algorithm [76] which recovers Kodaira's classification for elliptic pencils in terms of a resolution procedure as indicated.

Hence, viewing the Kodaira classification Thm. 1.6 of degenerate fibers in an elliptically fibered surface as an application of the classification of simple singularities, so far, I have attempted to discuss singularities, and more specifically simple singularities, in terms of their deformations and their resolutions. The ADE theme evolving from this discussion is well-known to be central: The ADE classification governs the classification of simple singularities as well as finite subgroups of $SU(2)$ in Thm. 1.2, the topology of their neighborhoods and thus their deformations in Thm. 1.4, and moreover the topology of their resolutions in Thm. 1.5. Another important occurrence of the ADE theme which I cannot leave unmentioned is the McKay correspondence:

Theorem 1.7. *Let $\widetilde{\Gamma} \subset SL(2,\mathbb{C})$ denote a non-trivial finite group. Then there exists a finite group $\Gamma \subset SU(2)$ of ADE type as in Thm.* 1.2 *which is conjugate to $\widetilde{\Gamma}$. Denote by $\mathfrak{g}$ the corresponding simply laced Lie algebra of the same ADE-type. Moreover, let ρ_i, $i \in \{0,\ldots,\mu\}$ denote the pairwise non-isomorphic irreducible representations of Γ, where ρ_0 is the trivial representation.*

1. *There is a one-to-one correspondence between the representations ρ_i, $i > 0$, and the simple roots of $\mathfrak{g}$. Moreover, if ρ denotes the natural two-dimensional representation of Γ on $\mathbb{C}^2$, then*

$$\rho_i \otimes \rho = \bigoplus_{j=0}^{\mu} a_{ij}\rho_j$$

defines the coefficients a_{ij} of a matrix A which obeys $A = 2I - \widetilde{C}$, where I denotes the $(\mu+1)\times(\mu+1)$ identity matrix and $\widetilde{C}$ denotes the Cartan matrix of the extended Dynkin diagram of $\mathfrak{g}$ [59].

2. *There exist locally free sheaves $\mathcal{R}_j$ on the minimal resolution $\widetilde{X}$ of $\mathbb{C}^2/\Gamma$ which outside the exceptional divisor are obtained from the equivariant bundle on $\mathbb{C}^2$ associated to the representation ρ_j. Moreover, if e_i denotes the irreducible component of the exceptional divisor which corresponds to the i^{th} simple root of $\mathfrak{g}$ by Thm.* 1.5*, then $c_1(\mathcal{R}_j)\cdot e_i = \delta_{ij}$ uniquely characterizes the first Chern class $c_1(\mathcal{R}_j)$ of each of these sheaves* [38, 53, 5].

There have been various successful attempts to generalize the classical McKay correspondence of Thm. 1.7 to higher dimensions. Let me particularly mention Ito and Reid's DUAL MCKAY CORRESPONDENCE [49], which for every finite subgroup $\Gamma \subset SL(n,\mathbb{C})$ states a one-to-one correspondence between the crepant divisors of a resolution of $\mathbb{C}^n/\Gamma$ and certain conjugacy classes in Γ, namely those of so-called JUNIOR elements. Moreover, in [11], Bridgeland, King and Reid show that for any nonsingular complex threefold Y and a finite automorphism group Γ of Y, the Hilbert scheme parametrizing Γ-clusters in Y gives a crepant resolution of Y/Γ, and that there is a derived equivalence between coherent sheaves on the resolution and coherent Γ sheaves on Y. A generalization of Thm. 1.7.2 to the three-dimensional case was found by Degeratu [22].

While Thm. 1.7 and its above-mentioned generalizations to higher dimensions state some of the best known classical incarnations of the McKay correspondence, it is probably not so well known that a similar correspondence holds much more generally for arbitrary quotient surface singularities:

Theorem 1.8. *Consider a finite subgroup $\Gamma \subset GL(2,\mathbb{C})$ which contains no pseudoreflections.*

1. *The quotient singularity $\mathbb{C}^2/\Gamma$ has only finitely many isomorphism classes of indecomposable reflexive modules* [46, 7, 34]*, and these are in one-to-one correspondence with the isomorphism classes of irreducible representations of the group Γ* [46].

2. *For every irreducible component e_i of the exceptional divisor in a minimal resolution $\pi\colon \widetilde{X} \to \mathbb{C}^2/\Gamma$, there exists a unique indecomposable reflexive module M_i. Denoting by $\widetilde{M_i}$ its locally free pull-back sheaf modulo torsion, the first Chern class of $\widetilde{M}_i$ is characterized by $c_1(\widetilde{M}_i)\cdot e_j = \delta_{ij}$ for all i, j* [86].

In other words, for general quotient surface singularities, there are SPECIAL REPRESENTATIONS which are in one-to-one correspondence with the irreducible components of the exceptional divisor of a minimal resolution. Ito, who has found a combinatorial criterion to determine special representations of cyclic groups [48], therefore calls the correspondence stated in Thm. 1.8 the SPECIAL

McKay correspondence. In the physics literature, this generalization of the classical McKay correspondence has already been used by Martinec, Moore and collaborators, see [58] along with the work built on that paper.

2. Some Background on Conformal Field Theory

An important lesson from the McKay correspondence points towards a deep relationship between representation theory (in this case, of finite groups) on the one hand and geometry (in this case, of the corresponding quotient singularities) on the other hand. Another link between representation theory and geometry is established by conformal field theory, although in most respects it is no less mysterious than the McKay correspondence. The basic ingredients of conformal field theories, their beautiful properties, as well as some results concerning their geometric interpretations are summarized in this section.

For completeness, let me state a rough definition of conformal field theory from my personal point of view. I apologize to the reader, since for lack of space, the details of this definition are left to [83]:

Definition 2.1. A (two-dimensional Euclidean) unitary conformal field theory with central charges c, $\overline{c}$ is given by

- a $\mathbb{C}$-vector space $\mathbb{H}$ with positive definite scalar product $\langle\cdot,\cdot\rangle$ and a compatible real structure $\phi \mapsto \phi^*$,
- a system $\langle\cdots\rangle$ of n-point functions, that is, a Poincaré covariant, local map

$$\mathbb{H}^{\otimes n} \to \mathrm{Maps}\left(\mathbb{C}^n - \bigcup_{i\neq j} D_{i,j}, \mathbb{C}\right), \quad D_{i,j} := \{(z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_i = z_j\}$$

  for every $n \in \mathbb{N}$, which is compatible with complex conjugation, and such that every function in the image is real analytic and allows an appropriate expansion about every partial diagonal D_{ij}.

The vector space $\mathbb{H}$ and the n-point functions must obey the following conditions:

1. $\mathbb{H}$ is a unitary representation of two commuting copies of a Virasoro algebra. The first, so-called left-handed Virasoro algebra is generated by $L_n,\ n \in \mathbb{Z}$, and a central element c, where

$$[L_n, L_m] = (m-n)L_{m+n} + \tfrac{c}{12}\delta_{n+m,0}m(m^2-1), \tag{2.1}$$

   and the second, right-handed one is generated by $\overline{L}_n,\ n \in \mathbb{Z}$, and $\overline{c}$ with the analogous commutator relations. Both Virasoro actions are

compatible with the real structure of $\mathbb{H}$. The central elements c, $\overline{c}$ act by multiplication with fixed constants, also denoted c, $\overline{c} \in \mathbb{R}$. The operators L_0 and $\overline{L}_0$ are self adjoint and positive semidefinite, and $\mathbb{H}$ decomposes into a direct sum of their simultaneous eigenspaces

$$\mathbb{H} = \bigoplus_{(h,\overline{h})\in R} \mathbb{H}_{h,\overline{h}}, \qquad \mathbb{H}_{h,\overline{h}} := \ker(L_0 - h) \cap \ker(\overline{L}_0 - \overline{h}).$$

That is, every vector in $\mathbb{H}$ is a sum of contributions from finitely many different eigenspaces $\mathbb{H}_{h,\overline{h}}$. Moreover, R does not have accumulation points and all $\mathbb{H}_{h,\overline{h}}$ are finite dimensional.

2. $\mathbb{H}$ possesses a unique vacuum Ω, that is $\mathbb{H}_{0,0} = \mathrm{span}_{\mathbb{C}}\{\Omega\}$ with $\Omega^* = \Omega$ and $\langle \Omega, \Omega \rangle = 1$.

3. The system $\langle \cdots \rangle$ of n-point functions is conformally covariant and represents an operator product expansion, and reflection positivity holds.

4. The partition function

$$Z(\tau) := \sum_{(h,\overline{h})\in R} \left(\dim_{\mathbb{C}} \mathbb{H}_{h,\overline{h}}\right) q^{h-c/24} \overline{q}^{\overline{h}-\overline{c}/24} = \mathrm{Tr}_{\mathbb{H}}\left(q^{h-c/24}\overline{q}^{\overline{h}-\overline{c}/24}\right)$$

with $\tau \in \mathbb{C}$, $\Im(\tau) > 0$, and $q := \exp(2\pi i \tau)$ is well defined for all values of τ in the complex upper halfplane, and it is invariant under modular transformations

$$\tau \mapsto \frac{a\tau + b}{c\tau + d}, \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z}).$$

5. The following universality condition holds: If $\mathbb{H} \subset \mathbb{H}'$ and $\langle \cdots \rangle'$ is a system of n-point functions on $\mathbb{H}'$, whose restriction to $\mathbb{H}$ gives $\langle \cdots \rangle$, and such that conditions 1.-4. hold for $\mathbb{H}'$ and $\langle \cdots \rangle'$, then $\mathbb{H} = \mathbb{H}'$.

From the above definition, representation theory and modular invariance have been built into conformal field theory by hand. The relation to geometry, however, is obscure. In fact, in many cases the precise relation to geometry is unknown. While so-called non-linear sigma model constructions, in physics, are believed to provide a map from certain geometries to conformal field theories, the mathematical details are far from understood. On the other hand, for certain classes of conformal field theories, geometric information can be extracted, in fact even geometric information related to the geometry of singularities reviewed in Section 1. This is particularly true if the space of states $\mathbb{H}$ of a conformal field theory carries an action of a left- and a right-handed SUPER-VIRASORO ALGEBRA:

Definition 2.2. The $N = 2$ SUPER-VIRASORO ALGEBRA with central charge c is the super-Lie algebra with even generators L_n, $n \in \mathbb{Z}$, J_n, $n \in \mathbb{Z}$, and a central element c, and odd generators G_r^+, G_r^-, where in the RAMOND (R) SECTOR all $r \in \mathbb{Z}$ and in the NEVEU-SCHWARZ (NS) SECTOR all $r \in \mathbb{Z} + \frac{1}{2}$. The L_n obey the Virasoro algebra (2.1), and furthermore, the following super-commutator relations hold:

$$\begin{aligned}
\left[L_n, G_r^\pm\right] &= (r - \tfrac{n}{2})G_{n+r}^\pm, \\
\left[G_r^+, G_s^-\right] &= 2L_{r+s} + (s-r)J_{r+s} + \tfrac{c}{3}(r^2 - \tfrac{1}{4})\delta_{r+s,0}, \quad \left[G_r^\pm, G_s^\pm\right] = 0, \\
[L_n, J_m] &= nJ_{m+n}, \quad \left[J_n, G_r^\pm\right] = \pm G_{n+r}^\pm, \quad [J_n, J_m] = \tfrac{c}{3}m\delta_{m+n,0}.
\end{aligned}$$

An $N = (2,2)$ SUPERCONFORMAL FIELD THEORY is a conformal field theory as in Def. 2.1 where both the left- and the right-handed Virasoro algebra are extended to commuting $N = 2$ super-Virasoro algebras. Moreover, $\mathbb{H}$ carries a compatible $\mathbb{Z}_2$ grading $\mathbb{H} = \mathbb{H}_b \oplus \mathbb{H}_f$, and the definition of locality is replaced by semi-locality, while the trace featuring in the definition of the partition function is only taken over the BOSONIC SUBSPACE $\mathbb{H}_b$. In general, $\mathbb{H}$ enjoys another compatible $\mathbb{Z}_2 \times \mathbb{Z}_2$ grading

$$\mathbb{H} = \mathbb{H}^{NS,NS} \oplus \mathbb{H}^{R,R} \oplus \mathbb{H}^{NS,R} \oplus \mathbb{H}^{R,NS}$$

where on $\mathbb{H}^{A,\overline{A}}$ the odd parts of the two $N = 2$ left- and right-handed Virasoro algebras, respectively, are represented in the A and the $\overline{A}$ sector, $A, \overline{A} \in \{R, NS\}$. Attention is very often restricted to so-called NON-CHIRAL SUPERCONFORMAL FIELD THEORIES, where the sectors $\mathbb{H}^{NS,R}$ and $\mathbb{H}^{R,NS}$ are trivial. Then, as a shorthand notation, one introduces $\mathbb{H}_k^A := \mathbb{H}^{A,A} \cap \mathbb{H}_k$ with $A \in \{R, NS\}$ and $k \in \{b, f\}$. For such theories one has

Definition 2.3. Consider a non-chiral $N = (2,2)$ superconformal field theory with central charges c, $\overline{c}$. Assume that the operator $J_0 - \overline{J}_0$ has only integral eigenvalues, such that the eigenvalues of $J_0 - \overline{J}_0$ are even on $\mathbb{H}_b$ and odd on $\mathbb{H}_f$. Then setting $q := \exp(2\pi i\tau)$ with $\tau \in \mathbb{C}$, $\Im(\tau) > 0$, and $y := \exp(2\pi iz)$ with $z \in \mathbb{C}$,

$$\mathcal{E}(\tau, z) := \mathrm{Str}_{\mathbb{H}^{R,R}}\left(y^{J_0} q^{h - c/24} \overline{q}^{\overline{h} - \overline{c}/24}\right)$$

is the CONFORMAL FIELD THEORETIC ELLIPTIC GENUS of the theory.

Using the properties of the $N = 2$ super-Virasoro algebra one shows that the conformal field theoretic elliptic genus only has nonzero contributions with $\overline{h} = \overline{c}/24$, it transforms covariantly under modular transformations, and it is invariant under smooth deformations of the underlying superconformal field theory into other $N = (2,2)$ superconformal field theories with the same central charges. In fact, for theories which are obtained by a non-linear sigma model construction from some Calabi-Yau variety Y, one expects that the conformal field theoretic elliptic genus agrees with the GEOMETRIC ELLIPTIC GENUS of Y. Vice versa, following [33, 62],

Definition 2.4. An $N = (2,2)$ superconformal field theory is said to be ASSOCIATED TO $K3$ if and only if the following conditions hold:

- For the left- and right-handed central charges, $c = \overline{c} = 6$ on $\mathbb{H}$.
- The theory is non-chiral, i.e. $\mathbb{H}^{NS,R} = \mathbb{H}^{R,NS} = \{0\}$, and the operators J_0 and $\overline{J}_0$ have integral eigenvalues only, where the eigenvalues of $J_0 - \overline{J}_0$ are even on $\mathbb{H}_b$ and odd on $\mathbb{H}_f$.
- The conformal field theoretic elliptic genus of the theory agrees with the geometric elliptic genus of a $K3$ surface,

$$\mathcal{E}(\tau, z) = 2\,\frac{\vartheta_2(\tau,z)^2\vartheta_3(\tau,0)^2\vartheta_4(\tau,0)^2 + \text{cycl.}}{\eta(\tau)^6}\,,$$

  where the $\vartheta_k(\tau, z)$ denote the classical Jacobi theta functions, and $\eta(\tau)$ is the Dedekind eta function.

Two such theories are equivalent, if and only if an isomorphism between the underlying $\mathbb{C}$ vector spaces exists which is compatible with the respective OPEs.[1]

All superconformal field theories which are associated to $K3$ in the sense of Def. 2.4 are believed to arise from non-linear sigma model constructions on $K3$ surfaces. Standard examples of such theories can be constructed by means of orbifolding: Starting from examples of so-called TOROIDAL $N = (2,2)$ SUPERCONFORMAL FIELD THEORIES with central charges $c = \overline{c} = 6$, which are well understood, orbifold techniques yield theories for which the conditions of Def. 2.4 are readily checked. For the underlying toroidal theories, in fact, geometric interpretations in terms of non-linear sigma models are mathematically well understood. Each such theory can be constructed as a non-linear sigma model on some flat torus T of complex dimension 2, equipped with a so-called B-FIELD, which is given by the de Rham cohomology class of a real closed two-form on T [16, 64].

The Definition 2.4 turns out to be sufficiently strong such that the moduli space of all theories associated to $K3$ can be determined, under the assumption that the standard deformation theory for them (see e.g. [27]) yields integrable deformations [17, 73, 6, 62]. Moreover, a partial completion of the smooth universal covering space of the moduli space can be identified with the parameter space of non-linear sigma models on $K3$ surfaces [6]:

Theorem 2.5. *Let $\mathbb{Z}^{4,20}$ denote the standard unimodular lattice of rank 24 and signature $(4,20)$ in $\mathbb{R}^{4,20} := \mathbb{Z}^{4,20} \otimes_{\mathbb{R}} \mathbb{R}$ with the compatible scalar product of signature $(4,20)$ and some chosen orientation. Let $\mathcal{T}^{4,20}$ denote the Grassmannian of maximal positive definite oriented subspaces of $\mathbb{R}^{4,20}$, which carries a natural action of $O^+(4,20;\mathbb{R})$, the group of those elements in $O(4,20;\mathbb{R})$*

[1] In particular, no $N = (2,2)$ marking of our superconformal field theory is fixed.

which preserve the orientation of such subspaces. By $\mathcal{T}_0^{4,20} \subset \mathcal{T}^{4,20}$ *we denote the set of all those oriented maximal positive definite subspaces* $x \subset \mathbb{R}^{4,20}$ *which have the property that* $x^\perp$ *does not contain any* ROOTS, *that is all* $\alpha \in x^\perp \cap \mathbb{Z}^{4,20}$ *obey* $\langle \alpha, \alpha \rangle \neq -2$. *Finally, with* $O^+(4,20;\mathbb{Z}) := O^+(4,20;\mathbb{R}) \cap O(\mathbb{Z}^{4,20})$ *and* $\mathcal{M}^{K3}$ *the so-called* MODULI SPACE OF SUPERCONFORMAL FIELD THEORIES ASSOCIATED TO $K3$,

$$\mathcal{M}^{K3} := O^+(4,20;\mathbb{Z}) \backslash \mathcal{T}_0^{4,20},$$

the following holds:

- *The partial completion* $\mathcal{T}^{4,20}$ *of the smooth universal covering space* $\mathcal{T}_0^{4,20}$ *of* $\mathcal{M}^{K3}$ *can be identified with the* PARAMETER SPACE OF NON-LINEAR SIGMA MODELS ON $K3$. *Namely, denoting by* X *the diffeomorphism type of a* $K3$ *surface,* $\mathcal{T}^{4,20}$ *is a cover of the space of pairs* (g, B) *where* g *denotes an Einstein metric on* X *and* B *is the de Rham cohomology class of a real closed two-form on* X, *a so-called* B-FIELD.

- *There is a one-to-one correspondence between the points of* $\mathcal{M}^{K3}$ *and superconformal field theories associated to* $K3$. *Smooth families of superconformal field theories associated to* $K3$ *are parametrized by smooth subvarieties of* $\mathcal{T}_0^{4,20}$.

Thm. 2.5 justifies the standard terminology according to which the specification of a superconformal field theory associated to $K3$ by means of the parameter space $\mathcal{T}_0^{4,0}$, that is, by means of a pair (g, B) with g an Einstein metric and B a B-field on $K3$, gives a GEOMETRIC INTERPRETATION of the theory. In fact, a non-linear sigma model construction on a $K3$ surface with Einstein metric g and B-field B is believed to yield a superconformal theory with geometric interpretation (g, B). Note that the so-called T-DUALITY GROUP $O^+(4,20;\mathbb{Z})$ acts by permuting an infinity of distinct geometric interpretations of a given superconformal field theory associated to $K3$.

As mentioned above, some standard examples of superconformal field theories associated to $K3$ can be obtained by orbifold constructions. On the other hand, some of the standard constructions of $K3$ surfaces are in fact orbifold constructions: Assume that a flat real four-torus T admits the choice of a complex structure such that $SU(2)$ acts naturally on the universal cover of T, and T enjoys a symmetry given by a non-trivial finite subgroup $\Gamma \subset SU(2)$. According to Fujiki [35], then Γ is cyclic of order $M \in \{2, 3, 4, 6\}$, or Γ is a binary dihedral group D_k of order $4(k-2)$ with $k = 4$ (and two inequivalent actions exist) or $k = 5$, or Γ is binary tetrahedral. In any case, the singularities of the quotient T/Γ are all of ADE type as in Thm. 1.2, and their minimal resolution of Thm. 1.5 is CREPANT. This means that there is a resolution $\pi\colon X \to T/\Gamma$ of all singularities, such that X is simply connected and admits an (orbifold limit of) a Ricci-flat Kähler metric which is induced from the complex structure and flat metric on T. In other words, X is a $K3$ surface. If Γ is cyclic, then T

equipped with its complex structure possesses an elliptic fibration with section such that the action of Γ is compatible with the fibration, and the resolution of T/Γ carries an induced elliptic fibration. These basic facts are used in the following theorem, which summarizes some of my own results, partly in joint work with W. Nahm, substantiating the geometric content of superconformal field theories associated to $K3$ in the case of orbifolds:

Theorem 2.6. *Consider an $N = (2, 2)$ superconformal field theory $\mathcal{C}$ associated to $K3$ which is obtained from a toroidal $N = (2, 2)$ superconformal field theory $\mathcal{T}$ by orbifolding by a finite group $\Gamma \subset SU(2)$. For the toroidal theory, assume that a geometric interpretation as non-linear sigma model on some complex torus T is given, such that the Γ action is induced by a symmetry of T. The geometric interpretation of $\mathcal{T}$ specifies the theory in terms of a Ricci-flat Kähler metric (that is, a flat metric) on T and a B-field, that is, a real de Rham cohomology class on T which is invariant under the induced Γ action.*

By X we denote the $K3$ surface which is obtained as a crepant resolution of the orbifold T/Γ. It carries an orbifold limit g of a Kähler-Einstein metric, induced by the chosen flat metric on T. Let $B_T \in H^2(X, \mathbb{R})$ denote the image of the B-field on the torus under the resolution procedure.

1. *The theory $\mathcal{C}$ possesses a geometric interpretation (g, B^{orb}), where*

$$|\Gamma| \cdot B^{orb} = \sqrt{|\Gamma|} \cdot B_T + \widehat{B}_\Gamma, \quad \widehat{B}_\Gamma \in H^2(X, \mathbb{Z}).$$

   The value of $\widehat{B}_\Gamma$ is given by a sum of contributions from the exceptional divisors over the quotient singularities $\mathcal{S} \subset T/\Gamma$ which is governed by the McKay correspondence Thm. 1.7.2 *as follows: For every quotient singularity $s \in \mathcal{S}$, whose local description is given by $0 \in \mathbb{C}^2/\Gamma_s$, $\Gamma_s \subset \Gamma$, let ρ_s denote the regular representation of Γ_s and $c_1(\mathcal{R}_s)$ the first Chern class of the locally free sheaf $\mathcal{R}_s$ on X associated to ρ_s by means of Thm.* 1.7.2. *Then* [81]

$$\widehat{B}_\Gamma = \sum_{s \in \mathcal{S}} c_1(\mathcal{R}_s).$$

2. *Assume that Γ is cyclic of order $M \in \{2, 3, 4, 6\}$, and that the torus T is elliptically fibered such that the Γ action is compatible with the fibration. Then fiberwise T-duality on the elliptic fibration of T induces fiberwise T-duality on X, see* [63] *for the explicit action on the degenerate fibers (which according to Thm.* 1.6 *are of Kodaira types I_0^* if $M = 2$, IV^* if $M = 3$, I_0^* and III^* if $M = 4$, and I_0^*, II^*, IV^* if $M = 6$). The resulting duality is a version of mirror symmetry on X* [63].

3. *If Γ is cyclic of order $M \in \{2, 3, 4, 6\}$ and the torus T is elliptically fibered such that the Γ action is compatible with the fibration, then in the superconformal field theory $\mathcal{C}$, the counterparts of the irreducible components of the exceptional divisor for the resolution of T/Γ can be determined explicitly: For the resolution of each singularity of T/Γ, they are given by*

discrete Fourier transforms of the corresponding twisted ground states in the orbifold conformal field theory $\mathcal{C}$ [63].

4. *If Γ is cyclic of order 4, then the superconformal field theory $\mathcal{C}$ also has a geometric interpretation (g, B) on a smooth algebraic $K3$ surface*

$$X(f_1, f_2): \quad f_1(x_0, x_1) + f_2(x_2, x_3) = 0 \text{ in } \mathbb{CP}^3,$$

where the f_k are homogeneous polynomials of degree 4, whose precise form depending on the geometric interpretation of $\mathcal{T}$ is stated in [82]. *Let ω_{FS} denote the class of the Kähler form associated to the Kähler metric on $X(f_1, f_2)$ which is induced by the Fubini-Study metric on $\mathbb{CP}^3$. Choose $\lambda \in \mathbb{R}^+$ such that with respect to the volume form $\frac{\lambda^2}{2}\omega_{FS} \wedge \omega_{FS}$, the surface $X(f_1, f_2)$ has volume $\frac{1}{2}$. Then our geometric interpretation (g, B) of $\mathcal{C}$ amounts to the Ricci-flat Kähler metric g on the $K3$ surface $X(f_1, f_2)$ in the class $\lambda\omega_{FS}$, and B-field $B = -\frac{1}{2}\omega_{FS}$* [82].

Note that in general, no explicit constructions are known for superconformal field theories associated to smooth $K3$ surfaces. However, Thm. 2.6.4 gives a simple ($\mathbb{Z}_4$ orbifold) construction for a class of superconformal field theories associated to smooth quartic $K3$ surfaces $X(f_1, f_2)$ with a fixed (natural) Kähler class and B-field.

3. Some Insights from Topological Field Theory

While in Section 2, I have tried to present a rather encouraging picture for superconformal field theories associated to $K3$, where geometry is manifestly visible within conformal field theory, the general situation is far more obscure. However, every $N = (2, 2)$ superconformal field theory possesses a "topological sector" which in many cases allows the extraction of geometric data. This section is devoted to the basic notions of these "topological sectors", in particular to the special geometries that arise from the study of families or moduli spaces of topological field theories.

For every $N = (2, 2)$ superconformal field theory the space of states $\mathbb{H}$ possesses interesting subspaces which carry the structure of FROBENIUS ALGEBRAS: Given such a conformal field theory, consider

$$\mathcal{A}^{c,c} := \left\{\phi \in \mathbb{H} \mid 2L_0\phi = J_0\phi,\ 2\overline{L}_0\phi = \ \ \overline{J}_0\phi\right\},$$
$$\mathcal{A}^{c,a} := \left\{\phi \in \mathbb{H} \mid 2L_0\phi = J_0\phi,\ 2\overline{L}_0\phi = -\overline{J}_0\phi\right\},$$

and $\mathcal{A}^{a,c}$, $\mathcal{A}^{a,a}$ are defined analogously. The $\mathcal{A}^{\bullet,\bullet}$ are often called the (CHIRAL, CHIRAL), (CHIRAL, ANTICHIRAL) RINGS, etc. [57], and indeed they inherit a ring structure from the operator product expansion mentioned in Def. 2.1.3, so each $\mathcal{A}^{\bullet,\bullet}$ is in fact an algebra. Additionally, there is a non-degenerate bilinear form on $\mathcal{A}^{\bullet,\bullet}$ inherited from the two-point functions on $\mathbb{H}$, and the axioms

of conformal field theory ensure that these structures conspire to those of a Frobenius algebra, as was first discovered by Witten [84, 85] and Dijkgraaf, Verlinde, Verlinde [26]:

Definition 3.1. A FROBENIUS ALGEBRA over $\mathbb{C}$ is a commutative associative $\mathbb{C}$ algebra with a unit e together with a non-degenerate bilinear form $\langle \cdot, \cdot \rangle \colon \mathcal{A} \times \mathcal{A} \to \mathbb{C}$, which is invariant, that is,

$$\forall\, a,\, b,\, c \in \mathcal{A}\colon \quad \langle a \cdot b, c\rangle = \langle a, b \cdot c\rangle.$$

Therefore, from now on I call the $\mathcal{A}^{\bullet,\bullet}$ the (CHIRAL, CHIRAL), (CHIRAL, ANTICHIRAL) ALGEBRAS, etc. In fact, the $\mathcal{A}^{\bullet,\bullet}$ are closely related to the structures that give rise to the elliptic genus of Def. 2.3: Under the conditions stated there, the space of states of an $N = (2,2)$ superconformal field theory enjoys a certain vector space isomorphism $\mathbb{H}^{NS,NS} \cong \mathbb{H}^{R,R}$ known as the SPECTRAL FLOW. Depending on the chosen normalizations and the chosen direction of the flow, under the spectral flow, those states which in the elliptic genus contribute to the leading order terms with $h = c/24$ and $\overline{h} = \overline{c}/24$ are mapped into one of the four algebras $\mathcal{A}^{\bullet,\bullet}$. Recalling that the conformal field theoretic elliptic genus is expected to agree with the geometric elliptic genus of a Calabi-Yau variety Y if our superconformal field theory is obtained by a non-linear sigma model construction from Y, and that the leading order terms of the geometric elliptic genus capture the Euler characteristic of Y, we expect to be able to identify the ring structure of $\mathcal{A}^{\bullet,\bullet}$ with a natural ring structure on $H^{*,*}(Y, \mathbb{C})$. This is one of the essential ideas of TOPOLOGICAL FIELD THEORY.

The Frobenius algebra $\mathcal{A}^{c,c}$ is particularly interesting for a superconformal field theory which has a Landau-Ginzburg model description (see for example [80] for an excellent introduction). A Landau-Ginzburg model describes the theory as a UV fixed point of the renormalization group flow for an $N = 2$ supersymmetric field theory with some superpotential $f(x)$, which in general is a quasihomogeneous polynomial in several variables $x = (x_0, \ldots, x_n)$. The (chiral, chiral) algebra then is given by the Jacobi algebra $\mathcal{J}$ for f as introduced in Def. 1.1.1. Indeed, by means of Landau-Ginzburg models there is a fundamental relation between certain $N = (2,2)$ superconformal field theories and singularity theory [78]. This relation is particularly beautiful in the case of the $N = (2,2)$ VIRASORO MINIMAL MODELS. These models arise from the classification of unitary lowest weight representations of the $N = 2$ super-Virasoro algebra. Indeed, if the central charge obeys $c < 3$, then by [10] such representations only exist at discrete values of c, namely for $c = 3k/(k+2)$ with $k \in \mathbb{N}$, and for each such value of the central charge only finitely many inequivalent unitary lowest weight representations exist. One obtains a well defined $N = (2,2)$ superconformal field theory at each central charge $c = \overline{c} = 3k/(k+2)$, where the Hilbert space $\mathbb{H}$ of states is the direct sum over all these representations, each taken in a two-fold "left-right symmetric" tensor product, one tensor factor for the action of the left handed

and the other for the right handed $N = 2$ super-Virasoro algebra. The resulting theory is the $N = (2, 2)$ VIRASORO MINIMAL MODEL OF TYPE A_{k+1}. Indeed, its (chiral, chiral) algebra agrees with the Jacobi algebra of the quotient singularity of type A_{k+1} of Thm. 1.2, and its (antichiral, chiral) algebra is trivial. Taking an appropriate sum over other left-right combinations of the available unitary lowest weight representations at given central charge one also obtains D and E type minimal models, where the (chiral, chiral) algebras accordingly agree with the Jacobi algebras of the corresponding simple surface singularities [78], and the (antichiral, chiral) algebras are trivial.

One thus comes to the conclusion that the Jacobi algebras of certain singularities must in fact be Frobenius algebras [77]. Indeed, this is well known [69, 70, 71] by means of the RESIDUAL PAIRING for the universal unfolding of the singularity. Furthermore, according to Def. 1.1.3, given such a semiuniversal unfolding F of f, at fixed t the function germ F_t gives a deformation of the Landau-Ginzburg potential f. The Landau-Ginzburg model with potential F_t is still a supersymmetric field theory, so the parameter space of such Landau-Ginzburg families obtains the structure of a Frobenius manifold:

Definition 3.2. A FROBENIUS MANIFOLD is a complex manifold M, such that the holomorphic tangent space $T_t^{1,0}M$ for every $t \in M$ is a Frobenius algebra over $\mathbb{C}$ with commutative associative multiplication $\cdot_t$, unit e_t, and non-degenerate quadratic form $g_t = \langle \cdot, \cdot \rangle_t$, obeying the following axioms:

- The quadratic form g defines a holomorphic metric on the holomorphic tangent bundle $T^{1,0}M$ with flat Levi-Civita connection ∇.
- The HIGGS FIELD $C\colon T^{1,0}M \to \Omega^1(M) \otimes T^{1,0}M$, $C_X Y := -X \cdot Y$ gives a smoth flat tensor field C, and the unit gives a smooth flat vector field e:
$$\nabla(C) = 0 \quad \text{and} \quad \nabla(e) = 0.$$
- There is a smooth vector field E, called the EULER FIELD, which obeys
$$\mathrm{Lie}_E(\cdot) = \cdot, \qquad \mathrm{Lie}_E(g) = (2 - d)\, g$$
for some $d \in \mathbb{C}$.

Indeed, an exact identification of the bases of semiuniversal unfoldings of the ADE singularities and the parameter spaces of their associated Landau-Ginzburg families, along with their Frobenius manifold structures, can be found in [8]. Moreover, the correlators in the corresponding topological field theories are holomorphic functions of the moduli, governed by differential equations which express the associativity of the operator product expansion mentioned in Def. 2.1.3 [25]. These differential equations exhibit a beautiful integrable structure [29].

The special geometry of Landau-Ginzburg families is in fact much richer, as was discovered by Cecotti [18, 19], and then generalized to arbitrary $N = 2$ supersymmetric quantum field theories by Cecotti and Vafa [20]. While the Frobenius manifold captures the holomorphic structure of spaces of $N = 2$ supersymmetric quantum field theories, the Hermitian metric, which for example in the superconformal case the $\mathcal{A}^{\bullet,\bullet}$ inherit from the space of states $\mathbb{H}$, involves the real structure of $\mathbb{H}$, in other words it involves anti-holomorphic parameters. In the physics literature, the process of taking into account these anti-holomorphic dependencies is dubbed the TOPOLOGICAL-ANTITOPOLOGICAL FUSION. In mathematical terms, these structures were first introduced by Dubrovin [30], following the work of Cecotti and Vafa. Namely, the holomorphic tangent bundle of a Frobenius manifold is equipped with a compatible flat real structure, and the so-called tt^* EQUATIONS are required to hold, imposing yet another flatness condition on the Higgs field of the Frobenius manifold. Dubrovin has reformulated these equations as a Riemann-Hilbert problem and has proved their integrability. Closing the circle, one again comes to the conclusion that for certain classes of singularities, the base of the semiuniversal unfolding should carry a tt^* geometry. Indeed, Hertling has shown that there is a canonical such structure on the semiuniversal unfolding of every hypersurface singularity [45]. Hertling also generalizes the tt^* geometries to so-called TERP STRUCTURES, using the language of twistor theory, and he argues that TERP structures offer a rich generalization of the notion of VARIATIONS OF HODGE STRUCTURES.

While the full beauty and impact of tt^* geometry and all its generalizations has remained an active and exciting area of research to the very day, a first application was almost immediately given by Cecotti and Vafa [21]. Using Dubrovin's integrability result [30] for the tt^* equations they were able to prove an ADE classification for a certain class of $N = (2, 2)$ superconformal field theories. Namely, consider an $N = (2, 2)$ superconformal field theory with Landau-Ginzburg description which admits non-degenerate massive deformations. Then the vanishing cycles in the C^∞-fibration (1.1), which arise from a semiuniversal unfolding F of the corresponding singularity, can be interpreted as "wave fronts of soliton solutions" close to a critical point of the potential F_t in the infrared regime of the Landau-Ginzburg model. In particular, the soliton spectrum in the infrared is given by the intersection theory of vanishing cycles or Lefschetz thimbles, while the superconformal field theory of interest is found in the UV limit. By Dubrovin's integrability result [30], the dimension of the Jacobi algebra and the intersection form of the vanishing cycles completely determine the solutions of the tt^* equations and thereby the geometry of the underlying moduli space. Moreover, Dubrovin's reformulation of the tt^* equations in terms of a Riemann-Hilbert problem reduces the classification problem for such theories to the classification of the relevant matrices which define the associated Riemann-Hilbert problem. In the case of $N = (2, 2)$ Virasoro minimal models with non-degenerate massive deformations, where it is also tacitly assumed that $J_0 - \overline{J}_0$ has only integral eigenvalues, the classification

turns out to reduce to the solution of the very same combinatorial problem which underlies the usual ADE classification, say, of simple singularities, as in Thm. 1.2:

Theorem 3.3. *There is an ADE classification of those $N = (2,2)$ Virasoro minimal models which possess non-degenerate massive deformations and which have only integral eigenvalues of $J_0 - \overline{J}_0$. In fact, for every such model the Frobenius algebra of (chiral, chiral) states is isomorphic to the Jacobi algebra of an ADE type surface singularity. Moreover, the moduli space of topological field theories obtained by massive deformations from such a model is isomorphic, as a tt^* geometry, to the base of a semiuniversal unfolding of the corresponding ADE singularity* [21].

Cecotti and Vafa's Thm. 3.3 provides a very satisfactory and explicit link between conformal field theory and singularity theory. In particular, the classical McKay correspondence of Thm. 1.7 is seen to play its role for those $N = (2,2)$ Virasoro minimal models which possess non-degenerate massive deformations and where $J_0 - \overline{J}_0$ only has integral eigenvalues, if we allow ourselves to identify the base of our semiuniversal unfolding with the Cartan algebra of the corresponding ADE Lie algebra, as explained in the discussion of Thm. 1.4.

However, the assumption that $J_0 - \overline{J}_0$ in our superconformal field theory possesses only integral eigenvalues amounts to a rather severe restriction. The existence of $N = (2,2)$ Virasoro minimal models which do not meet this assumption turns out to be very likely. In fact, Gannon has given a classification of numerous candidates for such models [37]: He has determined all modular invariant sums of left-right-handed combinations of characters of the unitary lowest weight representations for the $N = 2$ super-Virasoro algebra at any central charge $c < 3$, whose leading order term is $q^{-c/24}\overline{q}^{-\overline{c}/24}$, ensuring the uniqueness of the vacuum from Def. 2.1.2. Hence the partition functions of all $N = (2,2)$ Virasoro minimal models must belong to Gannon's list. The list at every central charge is finite but exhibits an abundance of additional candidates for partition functions, beyond the ADE classified ones of the models in Thm. 3.3. Using Gannon's list, Gray [40] was able to rederive the Cecotti-Vafa ADE classification from a purely conformal field theoretic point of view, by showing that the partition functions of the (ADE classified) $N = (2,2)$ Virasoro minimal models classified by Cecotti and Vafa agree precisely with those partition functions in Gannon's list for which the operator $J_0 - \overline{J}_0$ has only integral eigenvalues. This latter condition is equivalent to the assumption that the theory enjoys SPACETIME SUPERSYMMETRY. Moreover, Gray has shown that every modular invariant function in Gannon's list can be obtained from the partition function of an ADE model by an orbifold procedure. This implies

Theorem 3.4. *The partition function of every $N = (2,2)$ Virasoro minimal model agrees with that of an orbifold of one of the ADE classified minimal*

models with spacetime supersymmetry [40]. *All modular invariant functions in Gannon's list* [37] *can be obtained by such an orbifolding.*

Since the basic consistency conditions for the relevant orbifoldings have also been checked in [40], these results amount to overwhelming evidence for the expectation that every function in Gannon's list is in fact the partition function of a well defined $N = (2, 2)$ superconformal field theory. The zoo of necessary orbifoldings is listed explicitly in [40] and awaits its geometric interpretation [41]. It would be exciting if the role of the classical McKay correspondence (Thm. 1.7) in the ADE classification of models with spacetime supersymmetry (Thm. 3.3) could be taken by the special McKay correspondence (Thm. 1.8) in the full classification of $N = (2, 2)$ Virasoro minimal models which is expected to follow from Thm. 3.4.

4. Further Directions

In the previous sections, the discussion was mostly focussed around surface singularities and their geometry governing certain quantum field theories. The situation in higher dimensions, unfortunately, is far less clear, and in particular there is a lack of examples where more detailed explicit investigations could be performed. From my personal point of view, two classes of examples give reason for hope, because preliminary results are already at hand: For superconformal field theories, explicit constructions associated to so-called BORCEA-VOISIN THREEFOLDS can be carried out [50], and for string theories on elliptically fibered Calabi-Yau threefolds, investigations of the underlying geometry are on their way [24].

The Borcea-Voisin threefolds are examples of Calabi-Yau varieties. Their first detailed study was carried out independently by Borcea [9] and Voisin [79] in the context of mirror symmetry, since apart from few exceptions these threefolds come naturally in mirror pairs:

Definition 4.1. Consider a $K3$ surface X which allows an antisymplectic involution σ, and an elliptic curve E with the standard antisymplectic involution ι acting by multiplication by (-1) on a universal cover of E. Denote by $\mathbb{Z}_2$ the group of order two whose generator is given by the action of (σ, ι) on $X \times E$. The threefold which is obtained from $(X \times E)/\mathbb{Z}_2$ by minimal resolution of all singularities is called a BORCEA-VOISIN THREEFOLD.

One checks that all Borcea-Voisin threefolds Y are Calabi-Yau varieties with vanishing first Betti number. Moreover, closed formulas for all Hodge numbers of Y are known [9, 79], which depend on the precise complex geometry of the $K3$ surface X that enters in the construction. Here, the existence of an antisymplectic involution σ on X is a crucial restriction. Nikulin has shown that there are precisely 75 families $\mathcal{M}_{(r,a,\delta)}$ of $K3$ surfaces which admit such an antisymplectic involution [66]. The parameters (r, a, δ) for each of these

families in fact specify the action of σ: A $K3$ surface X is a member of the family $\mathcal{M}_{(r,a,\delta)}$ if and only if it admits an antisymplectic involution σ such that for the induced action of σ on $H^2(X,\mathbb{Z})$, the invariant sublattice $S \subset H^2(X,\mathbb{Z})$ has the following characteristic properties: The rank of S is $r \in \{0,\ldots,20\}$, its discriminant is $S^*/S \cong (\mathbb{Z}_2)^a$ with $a \in \{0,\ldots,11\}$, and finally, δ is the parity of the quadratic form $\langle\cdot,\cdot\rangle$ induced by the cup product on $H^*(X,\mathbb{Z})$. In other words, $\delta = 0$ if for all $s^* \in S^*$, $\langle s^*, s^*\rangle \in \mathbb{Z}$, and $\delta = 1$ otherwise. The parameters (r,a,δ) actually determine the isomorphism class of the lattice S uniquely [65], and except for the case $(r,a,\delta) = (11,11,1)$ its embedding into $H^2(X,\mathbb{Z})$ is unique up to automorphisms. Then the family $\mathcal{M}_{(r,a,\delta)}$ is the moduli space of S-polarized $K3$ surfaces [28]. The list of 75 possible three-tuples (r,a,δ) in [66] is almost symmetric under $(r,a,\delta) \leftrightarrow (20-r,a,\delta)$, and Borcea and Voisin have observed independently that every pair of Borcea-Voisin threefolds Y, $\check{Y}$ constructed from a pair of $K3$ surfaces X, $\check{X}$ with $X \in \mathcal{M}_{(r,a,\delta)}$ and $\check{X} \in \mathcal{M}_{(20-r,a,\delta)}$ is in fact a mirror pair [9, 79].

To construct superconformal field theories associated to Borcea-Voisin threefolds, one needs to take the tensor product $\mathcal{C}_X \otimes \mathcal{C}_E$ of a superconformal field theory $\mathcal{C}_X$ associated to a $K3$ surface X and a superconformal field theory $\mathcal{C}_E$ associated to an elliptic curve E, and then perform an orbifolding, induced by the geometric $\mathbb{Z}_2$ action of (σ,ι) on $X \times E$ as above. From Thm. 2.6, any orbifold conformal field theory associated to $K3$ is a good candidate for $\mathcal{C}_X$, while superconformal field theories $\mathcal{C}_E$ associated to elliptic curves are generally completely under control. However, a priori it is far from obvious whether any of the families $\mathcal{M}_{(r,a,\delta)}$ contain orbifold limits of $K3$. Moreover, compatibility of the induced B-field values (see Thm. 2.6.1) with the action of (σ,ι) is not obvious but a necessary precondition for a well defined orbifold to exist. In joint work with M. Khalid we are addressing these questions, and we already have preliminary results [50]:

Theorem 4.2. *Let $\mathcal{M}_{(r,a,\delta)}$ denote a family of $K3$ surfaces from Nikulin's classification* [66] *which allows an antisymplectic involution σ as above.*

1. *In the family $\mathcal{M}_{(r,a,\delta)}$, every $K3$ surface X is a $\mathbb{Z}_2$ orbifold limit of $K3$ if and only if $(r,a,\delta) = (18,4,0)$.*

2. *The family $\mathcal{M}_{(r,a,\delta)}$ contains a $K3$ surface X which is a $\mathbb{Z}_4$ orbifold limit of $K3$ if and only if $(r,a,\delta) = (20,2,1)$. The family $\mathcal{M}_{(20,2,1)}$ has only one element.*

3. *The family $\mathcal{M}_{(14,6,0)}$ contains a $K3$ surface X which is a $\mathbb{Z}_3$ orbifold limit of $K3$; the family $\mathcal{M}_{(18,4,1)}$ contains a $K3$ surface X which is a $\mathbb{Z}_6$ orbifold limit of $K3$.*

In 1.-3. above, the induced B-field for each of the corresponding orbifold conformal field theories associated to X is invariant under the induced action of the antisymplectic involution σ on X.

Even before the classification of all orbifold limits of $K3$ within the families $\mathcal{M}_{(r,a,\delta)}$ has been completed, it follows that we obtain explicit constructions for an interesting class of superconformal field theories from the study of Borcea-Voisin threefolds. In particular, those superconformal field theories deserve special attention which are associated to Borcea-Voisin threefolds whose underlying $K3$ surface has the complex structure of $X \in \mathcal{M}_{(20,2,1)}$: Using Thm. 2.6.4 it follows that these theories should have competing geometric interpretations on threefolds that arise as $\mathbb{Z}_2$ orbifolds from the product of certain quartic $K3$ surfaces with an elliptic curve. Hence these conformal field theories give a particularly interesting class of models for future investigations, allowing the construction of superconformal field theories associated to a family of threefolds for which there were no known constructions up to now [50].

Elliptically fibered Calabi-Yau threefolds with section provide another class of higher dimensional geometries, where interesting singularities occur naturally in a setting that is relevant for string theory. Indeed, these Calabi-Yau threefolds feature in a string-string duality, which declares that type IIA string theories or their F-theory limits on elliptically fibered Calabi-Yau threefolds with section are dual to $E_8 \times E_8$ heterotic strings on the product of an elliptic curve and a $K3$ surface. The mathematics of such string-string dualities is far from understood to date. In the context of this note, the most pressing questions concern highly degenerate fibers in elliptic fibrations:

Indeed, consider a (relatively minimally) elliptically fibered Calabi-Yau threefold $p\colon Y \to \Sigma$. This means [61, 42, 24] that Y and Σ are smooth, that all fibers of p are one dimensional and geometrically connected, and that Y contains no contractible surface whose contractible fibers lie in the fibers of p. Moreover, the generic fiber of p is an elliptic curve, the reduced discriminant locus is a divisor in Σ, and for every smooth curve $C \subset \Sigma$ which does not contain singular points of the reduced discriminant locus, the surface $p^{-1}(C)$ is smooth. We also assume that the fibration possesses a section. Let $\Delta \subset \Sigma$ denote the discriminant of the fibration, a divisor, which decomposes into irreducible components $\Delta = \cup_{i=0}^{D} \Delta_i$. Then our assumptions together with the Kodaira classification Thm. 1.6 ensure that for each $i \in \{0, \ldots, D\}$, the fibers of p over Δ_i generically have some fixed Kodaira type. However, the geometry of those fibers which occur over singular points of Δ, in particular over the intersections $\Delta_i \cap \Delta_j$ of two different components of Δ, is more complicated. Some beautiful work in this context has been done already, most importantly by Miranda [61] and by Grassi and Morrison [39]. However, a complete picture has not yet been established. In particular, the prediction from string-string dualities, which associates so-called CHARGED MATTER MULTIPLETS to such highly degenerate fibers, is still quite obscure from a mathematical point of view. In joint work with A. Degeratu we are addressing these issues; for example, [24] contains a refinement of Tate's algorithm which allows us to understand the specific degenerations of elliptic fibrations over non-trivial intersections $\Delta_i \cap \Delta_j$, $i \neq j$. We have already completed a detailed mathematical introduction to the topic [23].

References

[1] V.I. Arnol'd, *Integrals of rapidly oscillating functions and singularities of projections of Lagrangian manifolds*, Functional Anal. Appl. **6** (1972), 222–225.

[2] V.I. Arnol'd, *Normal forms for functions near degenerate critical points, the Weyl groups A_k, D_k, E_k and Lagrangian singularities*, Functional Anal. Appl. **6** (1972), 254–272.

[3] V.I. Arnol'd, S.M. Gusein-Zade, A.N. Varchenko, *Singularities of differentiable maps I, II*, Birkhäuser, Boston-Basel-Stuttgart, 1985.

[4] M. Artin, *On isolated rational singularities of surfaces*, Amer. J. Math. **88** (1966), 129–136.

[5] M. Artin, J.-L. Verdier, *Reflexive modules over rational double points*, Math. Ann. **270** (1985), 79–82.

[6] P. Aspinwall, D. Morrison, *String theory on $K3$ surfaces*, in: Mirror symmetry II, AMS/IP Stud. Adv. Math. **1**, Amer. Math. Soc., Providence, RI (1997), 703–716.

[7] M. Auslander, *Rational singularities and almost split sequences*, Trans. Amer. Math. Soc. **293** (1986), 511–531.

[8] B. Blok, A. Varchenko, *Topological conformal field theories and the flat coordinates*, Int. J. Mod. Phys. **A7** (1992), 1467–1490.

[9] C. Borcea, *K3 surfaces with involution and mirror pairs of Calabi-Yau manifolds*, in: Mirror symmetry II, AMS/IP Stud. Adv. Math. **1**, Amer. Math. Soc., Providence, RI (1997), 717–743.

[10] W. Boucher, D. Friedan, A. Kent, *Determinant formulae and unitarity for the $N = 2$ superconformal algebras in two dimensions or exact results on string compactification*, Phys. Lett. **B172** (1986), 316–322.

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

[12] E. Brieskorn, *Rationale Singularitäten komplexer Flächen*, Invent. Math. **4** (1968), 336–358.

[13] E. Brieskorn, *Die Monodromie der isolierten Singularitäten von Hyperflächen*, Manuscripta math. **2** (1970), 103–161.

[14] E. Brieskorn, *Singular elements of semi-simple algebraic groups*, in: Actes du Congrès Intern. Math. 1970, t.2, Gauthier-Villars, Paris (1971), 279–284.

[15] E. Brieskorn, *Singlaritäten*, Sonderdruck aus Jber. Deutsch. Math.-Verein. **78,** H.2 (1976), 93–112.

[16] A. Casher, F. Englert, H. Nicolai, A. Taormina, *Consistent superstrings as solutions of the $D = 26$ bosonic string theory*, Phys. Lett. **B162** (1985), 121–126.

[17] S. Cecotti, *$N = 2$ supergravity, type IIB superstrings, and algebraic geometry*, Commun. Math. Phys. **131** (1990), 517–536.

[18] S. Cecotti, *$N = 2$ Landau-Ginzburg vs. Calabi-Yau σ-models: Non-perturbative aspects*, Int. J. Mod. Phys. **A6** (1991), 1749–1813.

[19] S. Cecotti, *Geometry of $N = 2$ Landau-Ginzburg families*, Nucl. Phys. **B355** (1991), 755–775.

[20] S. Cecotti, C. Vafa, *Topological–anti-topological fusion*, Nucl. Phys. **B367** (1991), 359–461.

[21] S. Cecotti, C. Vafa, *On classification of $N = 2$ supersymmetric theories*, Commun. Math. Phys. **158** (1993), 569–644.

[22] A. Degeratu, *Geometrical McKay correspondence for isolated singularities*; preprint math.DG/0302068.

[23] A. Degeratu, K. Wendland, *Friendly giant meets pointlike instantons? On a new conjecture by John McKay*, to appear in "Moonshine – The first quarter century and beyond, a workshop on the moonshine conjectures and vertex algebras", LMS Lecture Notes Series.

[24] A. Degeratu, K. Wendland, *How intersection numbers are ruled by index theory*, in preparation.

[25] R. Dijkgraaf, H. Verlinde, E. Verlinde, *Notes on topological string theory and 2D gravity*, in: String theory and quantum gravity (Trieste, 1990), World Sci. Publ., River Edge, NJ (1991), 91–156.

[26] R. Dijkgraaf, H. Verlinde, E. Verlinde, *Topological strings in $d < 1$*, Nucl. Phys. **B352** (1991), 59–86.

[27] L.J. Dixon, *Some world-sheet properties of superstring compactifications, on orbifolds and otherwise*, in: Superstrings, unified theories and cosmology (Trieste, 1987), ICTP Ser. Theoret. Phys. **4**, World Sci. Publ., Teaneck, NJ (1988), 67–126.

[28] I.V. Dolgachev, *Mirror symmetry for lattice polarized $K3$ surfaces. Algebraic geometry, 4*, J. Math. Sci. **81** (1996), 2599–2630.

[29] B. Dubrovin, *Integrable systems in topological field theory*, Nucl. Phys. **B379** (1992), 627–689.

[30] B. Dubrovin, *Geometry and integrability of topological-antitopological fusion*, Commun. Math. Phys. **152** (1993), 539–564.

[31] A.H. Durfee, *Fibered knots and algebraic singularities*, Topology **13** (1974), 47–59.

[32] P. Du Val, *On isolated singularities of surfaces which do not affect the conditions of adjunction I*, Proc. Cambridge Phil. Soc. **30** (1934), 453–465.

[33] Tohru Eguchi, Hirosi Ooguri, A. Taormina, Sung-Kil Yang, *Superconformal algebras and string compactification on manifolds with $SU(n)$ holonomy*, Nucl. Phys. **B315** (1989), 193–221.

[34] H. Esnault, *Reflexive modules on quotient surface singularities*, J. reine angew. Math. **362** (1985), 63–71.

[35] Akira Fujiki, *Finite automorphism groups of complex tori of dimension two*, Publ. Res. Inst. Math. Sci. **24** (1988), 1–97.

[36] A.M. Gabriélov, *Intersection matrices of certain singularities*, Functional Anal. Appl. **7** (1973), 182–193.

[37] T. Gannon, $U(1)^m$ *modular invariants,* $N = 2$ *minimal models, and the quantum Hall effect*, Nucl. Phys. **B491** (1997), 659–688.

[38] G. Gonzalez-Sprinberg, J.-L. Verdier, *Construction géometrique de la correspondance de McKay*, Ann. Sci. École Norm. Sup. (4) **16** (1983), 409–449.

[39] A. Grassi, D. Morrison, *Group representations and the Euler characteristic of elliptically fibered Calabi-Yau threefolds*, J. Alg. Geom. **12** (2003), 321–356.

[40] O. Gray, *On the complete classification of unitary* $N = 2$ *minimal superconformal field theories*, University of Augsburg Doctoral Thesis, 2009.

[41] O. Gray, K. Wendland, *work in progress.*

[42] M. Gross, *A finiteness theorem for elliptic Calabi-Yau threefolds*, Duke Math. J. **74** (1994), 271–299.

[43] S.M. Gusein-Zade, *Intersection matrices for certain singularities of functions of two variables*, Functional Anal. Appl. **8** (1974), 10–13.

[44] S.M. Gusein-Zade, *Dynkin diagrams of the singularities of functions of two variables*, Functional Anal. Appl. **8** (1974), 295–300.

[45] C. Hertling, tt^* *geometry, Frobenius manifolds, their connections, and the construction for singularities*, J. reine angew. Math. **555** (2003), 77–161.

[46] J. Herzog, *Ringe mit nur endlich vielen Isomorphieklassen von maximalen, unzerlegbaren Cohen-Macaulay-Moduln*, Ann. Math. **233** (1978), 21–34.

[47] Heisuke Hironaka, *Resolution of singularities of an algebraic variety over a field of characteristic zero I, II*, Ann. of Math. (2) **79** (1964), 109–326.

[48] Yukari Ito, *Special McKay correspondence*, in: Geometry of toric varieties, Sém. Congr. **6**, Soc. Math. France, Paris (2002), 213–225.

[49] Yukari Ito, M. Reid, *The McKay correspondence for finite subgroups of* $SL(3, \mathbb{C})$, in: Higher-dimensional complex varieties (Trento, 1994), de Gruyter, Berlin 1996, 221–240.

[50] M. Khalid, K. Wendland, *SCFTs on higher dimensional cousins of* $K3$, in preparation.

[51] F. Klein, *Vorlesungen über das Ikosaeder und die Auflösung der Gleichung vom fünften Grade*, Leipzig 1884.

[52] F. Klein, *Über die hypergeometrische Reihe*, Vorlesungsausarbeitung, Göttingen 1894.

[53] H. Knörrer, *Group representations and the resolution of rational double points*, in: Finite groups – coming of age (Montreal, Que., 1982), Contemp. Math. **45**, Amer. Math. Soc., Providence, R.I., 1985, 175–222.

[54] Kunihiko Kodaira, *On compact complex analytic surfaces* I, Ann. Math. **71** (1960), 111–152; *On compact analytic surfaces* II, Ann. Math. **77** (1963), 563–626; III, Ann. Math. **78** (1963), 1–40.

[55] J. Kollár, *Resolution of singularities - Seattle lecture*, preprint arXiv:math/0508332 [math.AG].

[56] K. Lamotke, *Die Homologie isolierter Singularitäten*, Math. Z. **143** (1975), 27–44.

[57] W. Lerche, C. Vafa, N.P. Warner, *Chiral rings in $N = 2$ superconformal theories*, Nucl. Phys. **B324** (1989), 427–474.

[58] E.J. Martinec, G. Moore, *On decay of K-theory*, preprint hep-th/0212059.

[59] J. McKay, *Graphs, singularities, and finite groups*, in: "The Santa Cruz Conference on Finite Groups", Proc. Symp. Pure Math. **37** (1980), 183–186.

[60] J. Milnor, *Singular points of complex hypersurfaces*, Ann. Math. Studies **61**, Princeton University Press, 1968.

[61] R. Miranda, *Smooth models for elliptic threefolds*, in: The birational geometry of degenerations (Cambridge, Mass., 1981), Progr. Math. **29**, Birkhäuser Boston, Mass., 1983, 85–133.

[62] W. Nahm, K. Wendland, *A hiker's guide to $K3$. Aspects of $N = (4, 4)$ superconformal field theory with central charge $c = 6$*, Commun. Math. Phys. **216** (2001), 85–138.

[63] W. Nahm, K. Wendland, *Mirror symmetry on Kummer type $K3$ surfaces*, Commun. Math. Phys. **243** (2003), 557–582.

[64] K.S. Narain, *New heterotic string theories in uncompactified dimensions < 10*, Phys. Lett. **B169** (1986), 41–46.

[65] V.V. Nikulin, *Integral symmetric bilinear forms and some of their applications*, Math. USSR Isv. **14** (1980), 103–167.

[66] V.V. Nikulin, *Discrete reflection groups in Lobachevsky spaces and algebraic surfaces*, Proceedings of the ICM 1986, Amer. Math. Soc., Providence, RI (1987), 654–671.

[67] F. Pham, *Vanishing homologies and the n variable saddlepoint method*, in: Singularities, Proc. Symp. Pure Math. **40.2** (1983), 319–333.

[68] F. Pham, *La descente des cols par les onglets de Lefschetz, avec vues sur Gauss-Manin, Systèmes différentiels et singularités*, Astérisque **130** (1985), 11–47.

[69] Kyoji Saito, *Primitive forms for a universal unfolding of a function with an isolated critical point*, J. Fac. Sci. Univ. Tokyo, Sect. IA Math. **28** (1982), 775–792.

[70] Kyoji Saito, *Period mapping associated to a primitive form*, Publ. RIMS **19** (1983), 1231–1264.

[71] Morihiko Saito, *On the structure of Brieskorn lattices*, Ann. Inst. Fourier Grenoble **39** (1989), 27–72.

[72] H.A. Schwarz, *Über diejenigen Fälle, in welchen die Gaußische hypergeometrische Reihe eine algebraische Funktion ihres vierten Elementes darstellt*, J. reine angew. Math. **75** (1872), 292–335.

[73] N. Seiberg, *Observations on the moduli space of superconformal field theories*, Nucl. Phys. **B303** (1988), 286–304.

[74] P. Slodowy, *Simple singularities and simple algebraic groups*, Lecture Notes in Mathematics **815**, Springer-Verlag Berlin, 1980.

[75] T.A. Springer, *The unipotent variety of a semi-simple group*, in: Algebraic Geometry Internat. Colloq., Tata Inst. Fund. Res., Bombay 1968, Oxford University Press, London (1969), 373–391.

[76] J. Tate, *Algorithm for determining the type of singular fiber in an elliptic pencil*, in: "Modular functions of one variable, IV", Lecture Notes in Mathematics **476**, Springer-Verlag Berlin 1975, 33–52.

[77] C. Vafa, *Topological Landau-Ginzburg models*, Mod. Phys. Lett. **A6** (1991), 337–346.

[78] C. Vafa, N. Warner, *Catastrophes and the classification of conformal theories*, Phys. Lett. **B218** (1989), 51–58.

[79] C. Voisin, *Miroirs et involutions sur les surface $K3$*, Astérisque **218** (1993), 273–323.

[80] N. Warner, *$N = 2$ supersymmetric integrable models and topological field theories*, Trieste HEP Cosmol. 1992, 143–179.

[81] K. Wendland, *Consistency of orbifold conformal field theories on $K3$*, Adv. Theor. Math. Phys. **5** (2001), 429–456.

[82] K. Wendland, *A family of SCFTs hosting all "very attractive" relatives of the $(2)^4$ Gepner model*, JHEP **0603** (2006), 102–150.

[83] K. Wendland, *Conformal field theory for mathematicians*, book manuscript currently under revision and to be completed soon.

[84] E. Witten, *On the structure of the topological phase of two-dimensional gravity*, Nucl. Phys. **B340** (1990), 281–332.

[85] E. Witten, *Two-dimensional gravity and intersection theory on moduli space*, in: Surveys in differential geometry (Cambridge, MA, 1990), Lehigh Univ. Bethlehem, PA (1991), 243–310.

[86] J. Wunram, *Reflexive modules on quotient surface singularities*, Math. Ann. **279** (1988), 583–598.

Author Index*

(Volumes II, III, and IV)

*Names of invited speakers only are shown in the Index.